Einführung von Begriff und Quantifizierung der Messunsicherheit auf Bayes-statistischer Grundlage in die psychologische Praxis am Beispiel der Personalauswahl und Potenzialbeurteilung durch Assessment Center

Von der Gemeinsamen Naturwissenschaftlichen Fakultät
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig
zur Erlangung des Grades einer
Doktorin der Naturwissenschaften
(Dr. rer. nat.)
genehmigte
D i s s e r t a t i o n

von Antonia Weise
aus Braunschweig

1. Referent: Prof. Dr. Heiner Erke,
Technische Universität Carolo-Wilhelmina zu Braunschweig
2. Referent: Prof. Dr. Dieter Frey,
Ludwig-Maximilians-Universität München

eingereicht am: 17. Juni 2002
mündliche Prüfung (Disputation) am: 23. Oktober 2002
Druckjahr: 2002

Antonia Weise
Email: WeiseAnton@aol.com

Zusammenfassung

Die mit Urteils- und Entscheidungsprozessen aufgrund unzureichender Information immer verbundene Unsicherheit bildet ein ernstes Problem hinsichtlich der Qualität der Beurteilung oder des Risikos einer Fehlentscheidung. Das gilt exemplarisch für die Personalauswahl und Potenzialbeurteilung durch Assessment Center (AC). Um die Unsicherheit zu quantifizieren, wird der in der physikalischen Messtechnik gebräuchliche, international genormte Begriff der Messunsicherheit interdisziplinär auf die psychologische Praxis übertragen. Die Messunsicherheit fußt auf der Bayes'schen Statistik und drückt den Kenntnismangel einer physikalischen Messgröße und damit die Genauigkeit einer Messung oder die Qualität eines Messergebnisses quantitativ aus. Bei der Übertragung werden zu bewertende Merkmale von Personen oder Alternativen in Analogie zu physikalischen Messgrößen betrachtet. Daraus folgt ein Auswerteverfahren für die Daten aus einem Urteils- oder Entscheidungsprozess, z.B. einem aktuellen AC, das EDV-Anwendung erlaubt und zu den Ergebnissen der Merkmalsbewertungen jeweils auch die Unsicherheiten liefert, was den kritischen Vergleich unterschiedlicher Ergebnisse zu einem Merkmal erleichtert. Außerdem lassen sich bei genügend umfangreichem Datenmaterial Unsicherheitskennwerte berechnen, die konventionell ermittelte Korrelationskoeffizienten zur Beurteilung der Konstruktvalidität eines AC ergänzen können. Das Auswerteverfahren wird anhand von drei AC-Serien mit vielen Teilnehmern evaluiert.

Abstract

Introduction of the concept and quantification of measurement uncertainty based on Bayesian statistics to psychological practice, exemplified by personnel selection and potential rating through assessment centers

The uncertainty, always connected with assessment and decision processes due to insufficient information, is a severe problem with regard to the quality of assessment or the risk of a wrong decision. This applies particularly to personnel selection and potential rating through assessment centers (AC). To quantify the uncertainty, the concept of measurement uncertainty, commonly used in physical metrology and internationally standardized, is interdisciplinarily transferred to psychological practice. The measurement uncertainty is based on Bayesian statistics and quantitatively expresses the lack of knowledge of a physical measurand and, thus, the accuracy of a measurement or the quality of a measurement result. The transfer is performed by treating attributes to be rated of persons or alternatives by analogy to physical measurands. This results in a procedure for analyzing the data obtained from an assessment or decision process, for instance, a current AC. This procedure enables electronic data processing to be applied and provides not only the results of attribute rating but also the uncertainties associated with them. These facilitate a critical comparison of different results of an attribute. Moreover, if sufficient data material is available, characteristic uncertainty values can be calculated for assessing the construct validity of an AC in addition to correlation coefficients obtained in a conventional way. The analyzing procedure is evaluated using three AC series with many participants.

Vorwort

In meiner beruflichen Tätigkeit als Psychologin im Bereich der Personalauswahl und Personalentwicklung in Unternehmen war ich häufig an Urteils- und Entscheidungsprozessen beteiligt, z.B. bei Einzel- oder Gruppenentscheidungen oder bei der Personalauswahl und Potenzialbeurteilung durch Assessment Center (AC). Dabei habe ich die bei der Bewertung der Entscheidungsalternativen bzw. AC-Teilnehmer durch unzureichende Information bewirkte Unsicherheit, die die Beurteilungsqualität mindert und das Entscheidungsrisiko erhöht, aber doch meist unbeachtet bleibt, oft als sehr ernstes Problem und methodischen Mangel empfunden. Dies umso mehr wegen meiner Kenntnis von den erfolgreichen internationalen wissenschaftlichen Bemühungen um das aus allgemeinerer Sicht ähnliche Problem der Messunsicherheit beim physikalischen Messen, ein langjähriges Arbeitsgebiet meines Vaters, Herrn Prof. Dr. Klaus Weise, an der Physikalisch-Technischen Bundesanstalt in Braunschweig. Daraus wurde mit der Frage: „Könnte man nicht zur Lösung des Problems der Unsicherheit in der psychologischen Praxis genau so oder ähnlich vorgehen wie in der Physik?“ schließlich die Idee zu dieser Arbeit geboren.

Durch zwei wesentliche Ansätze wurde es möglich, diese Frage zu bejahen. Der erste Ansatz betrifft den theoretischen, prinzipiellen Aspekt der Frage, der zweite Ansatz den Aspekt der Durchführung in der Praxis. Der erste Ansatz besteht in der begrifflichen Identifizierung eines im Urteils- oder Entscheidungsprozess zu bewertenden Merkmals, z.B. einer wichtigen Eigenschaft einer Person, mit einer physikalischen Messgröße, einer zu messenden Eigenschaft eines Messobjekts. Aufgrund dieses Ansatzes ist es bereits im Prinzip möglich, alle in der Physik erarbeiteten Konzepte, Methoden und Verfahren zur Messunsicherheit unmittelbar auf Urteils- und Entscheidungsprozesse in der Psychologie zu übertragen. Der zweite Ansatz beruht auf einer neuen, besonderen Art der Bewertung eines Merkmals, wodurch das Kernproblem des zweiten Aspekts gelöst werden konnte, nämlich die möglichst einfache Erfassung von Information für die Quantifizierung der Unsicherheit im laufenden Prozess, z.B. in einem aktuellen Assessment Center. Bei diesem Ansatz stehen einem geringen Mehraufwand für die Datenerfassung der Vorteil der Quantifizierung der Unsicherheit und die Möglichkeit, Computer zur schnellen Auswertung und Unterstützung einzusetzen, gegenüber.

Den Referenten dieser Arbeit, Herrn Prof. Dr. Heiner Erke, Technische Universität Carolo-Wilhelmina zu Braunschweig, und Herrn Prof. Dr. Dieter Frey, Ludwig-Maximilians-Universität München, möchte ich für die engagierte und aktive Betreuung, Aufmunterung, Anregungen und konstruktive Kritik danken, ebenso meinem Vater,

auch für viele Diskussionen und Hinweise zum Thema Messunsicherheit. Ich danke weiterhin dem Niedersächsischen Justizministerium in Hannover, der Justizvollzugsschule des Landes Niedersachsen in Wolfenbüttel und der Stinnes AG Holding in Mülheim (Ruhr) für die freundliche Überlassung der anonymisierten Daten aus durchgeführten Assessment Centern, die ich für die Evaluierung des in dieser wissenschaftlichen Arbeit beschriebenen Auswerteverfahrens benutzt habe. Ebenfalls danke ich der auf dem Gebiet der Personalentwicklung (PE) tätigen Firma PE-Solution in Burgdorf bei Hannover für die großzügige Hilfe und Erstellung der Programme auf der Grundlage dieser Arbeit für die Evaluierung und Anwendung des Verfahrens. Die Bereitwilligkeit der Justizvollzugsschule, die von mir vorgeschlagene neue Art der Bewertung in die Konzeption ihres Assessment Centers aufzunehmen, was die Evaluierung des Verfahrens sehr erleichtert hat, möchte ich noch besonders hervorheben.

Essen, im Juni 2002

Antonia Weise

Inhaltsverzeichnis

1 Einführung

1.1 Personalauswahl und Potenzialbeurteilung durch Assessment Center

Personalauswahl und Potenzialbeurteilung gehören zu den wichtigsten Aufgaben der Arbeits-, Betriebs- und Organisationspsychologie im Bereich des Personalmanagements in Unternehmen z.B. bei Einstellungen, Beförderungen oder Funktionsübertragungen. Denn für den Unternehmenserfolg ist es entscheidend, dass bestgeeignete Kräfte im Hinblick auf die jeweils zu lösenden Aufgaben, Stellenanforderungen und vorhandenen Mittel eingesetzt werden. Personelle Fehlentscheidungen können zu empfindlichen Einbußen führen (Obermann, 1992; Schmidt und Hunter, 2000). Das gilt auch ganz allgemein für komplexe und schwerwiegende Entscheidungen zwischen Alternativen, z.B. in der Unternehmenspolitik oder bei der Beschaffung hochwertiger Einrichtungen. Das Risiko von Fehlentscheidungen kann durch Anwendung bewährter psychologischer bzw. betriebswirtschaftlicher Methoden im Urteils- oder Entscheidungsprozess vermindert werden. Der oft erhebliche Aufwand dieser Methoden lässt sich mit Hilfe der elektronischen Datenverarbeitung (EDV) reduzieren. Auch kann damit die Wirksamkeit der Methoden durch konsequente Auswertung jeglicher vorhandenen Information und strikte Beachtung aller Anforderungen und Bedingungen verbessert werden. Dazu soll diese Arbeit beitragen, insbesondere durch Quantifizierung der mit der Bewertung einer bestimmten Person oder Alternative verbundenen Unsicherheit und damit des Risikos, sich für diese Person oder Alternative zu entscheiden.

Ein mit dieser Arbeit verwandtes Themengebiet stellt die Entscheidungsforschung dar (Gigerenzer und Selten, 2001; Jungermann, Pfister und Fischer, 1998). Sie führt Erkenntnisse aus unterschiedlichen wissenschaftlichen Disziplinen zusammen, unter anderem aus der Betriebswirtschaft, der es um die Optimierung wirtschaftlicher Entscheidungsprozesse geht (Eisenführ und Weber, 1994), und aus der Statistik, die Methoden zum Umgang mit Hypothesen und Daten auf mathematischer und wahrscheinlichkeitstheoretischer Grundlage beiträgt. Die psychologische Entscheidungsforschung unterscheidet zwei Richtungen, die präskriptive und die deskriptive Entscheidungsforschung. Die präskriptive Entscheidungsforschung setzt rationales Denken beim Entscheidenden voraus und entwickelt als Hilfen für die Entscheidungsfindung formalisierte Verfahren zur Strukturierung und Verarbeitung von Information. Hierzu zählen auch Unterstützungssysteme auf entscheidungstheoretischer Basis wie das in dieser Arbeit vorgestellte Verfahren. Ein Überblick über Entscheidungshilfetechnologien findet

sich bei Zimolong und Rohrmann (1988). Zur Unterstützung der Entscheidungsfindung erstellte Computerprogramme sind bei Nabe und Schmid (1997) sowie Nagel (1993) dargestellt. Die deskriptive Entscheidungsforschung beschäftigt sich mit der Beschreibung des tatsächlichen menschlichen Entscheidungsverhaltens, das aufgrund der beschränkten kognitiven Verarbeitungskapazität häufig von rein rationalen Modellen und Vorhersagen abweicht (Gigerenzer, Todd and ABC Research Group, 1999). Die Zusammenführung beider Ansätze kann zur Verbesserung von Entscheidungen und deren Vorhersage beitragen.

Angewendete Methoden zur Unterstützung der Entscheidungsfindung bei der Personalauswahl und Potenzialbeurteilung sind das Assessment Center (AC) und das Einstellungsinterview (Kleinmann, 1997; Lang–von Wins und von Rosenstiel, 2000). Das AC ergänzt andere Verfahren, wie z.B. die Sichtung von Bewerbungsunterlagen, um die Verhaltenskomponente. Im AC werden Teilnehmer (Kandidaten) gemeinsam oder getrennt verschiedenen praxisnahen Übungen unterzogen. Dabei bewerten Beobachter mehrere für die jeweiligen Anforderungen wichtige Merkmale (Eigenschaften) der Teilnehmer anhand des beobachteten Verhaltens der Teilnehmer. Die Bewertungen werden häufig auf einer Skala vorgenommen und mit Beschreibungen des Verhaltens ergänzt. In der anschließenden Beobachterkonferenz werden die Bewertungen harmonisiert, d.h. die Beobachter diskutieren, auf welche gemeinsame Bewertung für jedes Merkmal eines Teilnehmers sie sich auf der Skala anhand ihrer Beobachtungen und ihrer eigenen Bewertungen einigen können. Dazu wird oft als Diskussionsgrundlage ein Mittelwert der einzelnen Bewertungen gebildet und die Beobachtungen werden dagegengestellt. Manchmal wird auch die Streuung der einzelnen Bewertungen ermittelt. Abschließend wird für jeden Teilnehmer ein Profil der gemeinsamen Bewertungen der einzelnen Merkmale festgestellt und gegebenenfalls eine Personalentscheidung getroffen.

Das AC ist in der Unternehmenspraxis zwar durchaus etabliert und erfährt bei Teilnehmern hohe Akzeptanz, jedoch überwiegt bei Einstellungen das weniger aufwändige Interview, obwohl dieses eine geringere prognostische Validität als das AC aufweist und dadurch das Risiko von Fehleinschätzungen dabei höher ist (Lang–von Wins und von Rosenstiel, 2000; Sarges, 1995). Manchmal wird auch ein gemischtes Verfahren angewendet, z.B. ein strukturiertes Interview (Jetter, 1996) mit mehreren Beobachtern und situativen Übungen wie ein Fachvortrag jedes Teilnehmers. Obwohl ein solches Verfahren nicht ausdrücklich AC genannt wird, kann es doch in verallgemeinernder Sicht als AC aufgefasst werden. Interview und Vortrag können einfach als Übungen eines AC oder ein Einstellungsgespräch allein als ein minimales AC mit nur einer einzigen Übung und nur einem Beobachter angesehen werden (Abschnitt 2.2.1). Deshalb wird in dieser Arbeit nicht grundsätzlich zwischen AC, Interview und ähnlichen Methoden

unterschieden, zumal das hier vorgestellte Verfahren der Erfassung und Auswertung von Information auf jene Methoden gleichermaßen angewendet werden kann.

Tatsächlich sind alle Bewertungen mehr oder weniger unsicher. Das braucht aber keineswegs zu bedeuten, dass man auf Anwendung der EDV und Quantifizierung statistischer Kennwerte verzichten muss. Es wird in dieser Arbeit gezeigt, dass zusammen mit der Auswertung von Information auch die damit verbundene Unsicherheit quantitativ erfasst und berücksichtigt werden kann. Es wird dazu interdisziplinär eine Anleihe aufgenommen bei der physikalischen Messtechnik oder allgemeiner bei der Metrologie, der Lehre vom Messen, in der der Begriff der Messunsicherheit eine große Rolle spielt (Abschnitte 1.3 und 3.3). Die Quantifizierung der Unsicherheit zu einer Bewertung oder allgemein zu einer Aussage kann anschaulich deren Genauigkeit oder Qualität ausdrücken oder das Vertrauen darin begründen.

Es soll dementsprechend in dieser Arbeit ein Auswerteverfahren für die Information aus einem einzelnen AC entwickelt werden, das die Anwendung der EDV erlaubt und zu der Bewertung der Merkmale eines Teilnehmers insbesondere auch quantitativ die Unsicherheit liefert und damit das Risiko einer gegebenenfalls zu treffenden Entscheidung zum Ausdruck bringt.

Die EDV findet derzeit Anwendung in einigen Bereichen des AC. Dabei lässt sich ein Trend zu computergestützten Bausteinen feststellen (Lehment, 1999). Als Beispiele sind hier bewährte Leistungs- und Persönlichkeitstests zu nennen, die auch in Computerversionen von einigen Firmen angeboten werden. Ein Überblick über im Personalmanagement eingesetzte Persönlichkeitstests findet sich bei Hossiep, Paschen und Mühlhaus (2000), über Intelligenzverfahren bei Daumenlang (1995). Üblich sind auch computersimulierte Übungen, Planspiele und Szenarien, wie praxisnahe Postkorb-Übungen bis hin zu Simulationen kompletter Unternehmen, z.B. mit dem Ziel, vernetztes Denken, Problemlösefähigkeit und Entscheidungsverhalten von Teilnehmern zu messen (Funke, 1993, 1995; Obermann, 1995; Strauß und Kleinmann, 1996). Beispiele dafür sind die Computerprogramme DISKO (Funke, 1991) und *Textilfabrik* (Hasselmann und Strauß, 1995). Weiterhin gibt es das *Interview-PC-System*, ein umfangreiches PC-gestütztes Expertensystem als Hilfe für die Vorbereitung und Durchführung strukturierter Interviews (Jetter, 1996), das auch zur Unterstützung eines AC, z.B. bei der Konstruktion, benutzt werden kann. Für diesen Zweck bietet es Datenbanken von Merkmalen und Verhaltensweisen sowie von darauf abzielenden Fragen für die Erstellung von Anforderungsprofilen und Interviewleitfäden. Es erlaubt u.a. auch die Auswertung sowie die graphische Darstellung von Bewertungsprofilen der Teilnehmer. Ähnliches leistet das Programmsystem ELIGO von der Firma ELIGO Psychologische

Personalsoftware, das für die Vorauswahl von Bewerbern konzipiert ist. Es unterstützt die Erstellung von Anforderungsprofilen, stellt passende computergestützte Tests zusammen und wertet diese aus. Im laufenden AC benutzt manche Unternehmensberatung zur Unterstützung der Beobachterkonferenz das bekannte Programm Excel von der Firma Microsoft für die Eingabe der Bewertungen und die Bildung und Darstellung von Diagrammen, z.B. Bewertungsprofilen der Teilnehmer. In der Forschung ist die EDV bei der Evaluierung des AC unabdingbar. Hier werden für die statistische Auswertung großer Datenmengen z.B. für die Validitäts-, Korrelations- und Faktorenanalyse die Programmsysteme SPSS und LISREL von der Firma SPSS eingesetzt (Kleinmann, 1997; Jöreskog und Sörbom, 1989). EDV-Anwendungen zum AC, in denen die Unsicherheit irgendeine Rolle spielt, sind allerdings nicht bekannt.

Einen guten Überblick über den allgemeinen Stand der Forschung zum AC gibt Kleinmann (1997). Hinweise zur Konstruktion eines AC werden in Abschnitt 4.1 gegeben.

1.2 Mess- und Entscheidungsaufgaben

Bei einer Messaufgabe in der Physik oder Technik liegt bei näherer, analoger und verallgemeinernder Betrachtung eine zum AC sehr ähnliche Situation vor: In unterschiedlichen Experimenten (Übungen) werden Messgrößen (Merkmale) von gleichartigen Messobjekten (Teilnehmern) unter Anwendung von Messinstrumenten (Beobachtern) gemessen. Die Messinstrumente liefern Messwerte (Bewertungen), diese Daten sind zu Messergebnissen auszuwerten. Ein Messergebnis ist ein Schätzwert für den zum Messobjekt gehörenden und zu ermittelnden wahren Wert einer Messgröße. Gegebenenfalls wird anschließend geprüft, ob das Messergebnis ein Soll erfüllt und danach eine Entscheidung getroffen, ob das Messobjekt, z.B. ein gefertigtes Exemplar eines industriellen Produkts, akzeptiert oder verworfen wird. (In Klammern stehen in diesem Abschnitt die jeweils analogen Begriffe im AC, die in Abschnitt 2.2 definiert werden.)

Auch in einem ganz allgemeinen Entscheidungsprozess (z.B. Kepner-Tregoe, 1971), z.B. des betrieblichen Managements, besteht eine zum AC analoge Situation: Mittels unterschiedlicher Untersuchungen (Übungen) wird von Entscheidungsträgern (Beobachtern) bewertet, in welchem Maße Eigenschaften (Merkmale) von in Betracht gezogenen gleichartigen Alternativen (Teilnehmern) gestellten Anforderungen genügen. Danach wird dann die Entscheidung für eine der Alternativen gefällt.

1.3 Messunsicherheit

Trotz sorgfältigster Messung oder Ermittlung von Information ist diese kaum je vollständig. Das gilt nicht nur für ein AC, sondern auch für jede Mess- und Entscheidungsaufgabe. Durch diesen Mangel an Information verbleibt immer eine Unsicherheit darüber, inwieweit neben dem aus der Auswertung der vorliegenden Information ermittelten Schätzwert für eine Messgröße oder für ein anderes Merkmal eines Messobjekts auch andere Schätzwerte noch vernünftig sind. Will man nun das Vertrauen in einen Schätzwert oder dessen Qualität ausdrücken, so gehört zu einem Messergebnis als Schätzwert für eine Messgröße immer auch die Messunsicherheit, die den Mangel in der Kenntnis der Messgröße quantifiziert.

Der aus der physikalischen Messtechnik (Metrologie) stammende Begriff der Messunsicherheit – nicht zu verwechseln mit dem „Messfehler“ – ist relativ jung. In den letzten Jahren wurden die Messunsicherheit auf der Bayes'schen Statistik begründet (Weise und Wöger, 1992, 1999) sowie allgemein anzuwendende Regeln zu ihrer Quantifizierung und Handhabung bei der Auswertung von Messungen international im ISO *Guide to the Expression of Uncertainty in Measurement* (GUM, 1993; auch als Europäische Vornorm ENV 13005, 1999) und national in den Normen DIN 1319-3 (1995) und DIN 1319-4 (1999) festgelegt. Die *Messunsicherheit* – auch kürzer *Unsicherheit* genannt – ist ein Maß für die unvollständige Kenntnis einer Messgröße und damit der Genauigkeit oder Qualität einer Messung. Sie kann das Vertrauen in ein Messergebnis für diese Messgröße vertiefen, gerade auch bei statistisch unzureichenden und nicht statistisch erhobenen Daten und bei sich nicht zufällig verhaltenen unsicheren Einflüssen wie sie in der Regel sowohl beim physikalischen Messen als auch im Interview und im AC vorkommen. Die in diesem Absatz erwähnte Literatur zur Messunsicherheit wird im Folgenden zusammenfassend mittels des Kürzels [Uns] zitiert.

Die Messunsicherheit ersetzt die *Messabweichung*, den früher so genannten „Messfehler“, als Maß für die Genauigkeit einer Messung. Sie darf mit diesem keinesfalls verwechselt werden. Denn Messunsicherheit ist keine andere Benennung für Messfehler, sondern umfasst einen ganz anderen Begriff. Die Messunsicherheit beschreibt quantitativ den Mangel in der Kenntnis einer Messgröße oder eines Merkmals und lässt sich aus der gewonnenen Information berechnen. Der Messfehler dagegen ist die Abweichung einer Bewertung von dem entsprechenden wahren Wert der Messgröße bzw. der wahren Ausprägung des Merkmals. Über diese Abweichung liegt aber nicht mehr Information vor als über die zu ermittelnde Messgröße bzw. das zu ermittelnde Merkmal selbst, d.h. in der Regel nur unvollständige Information. Deshalb wird heute in der Messtechnik der Messfehler für die Charakterisierung der Genauigkeit einer Messung

als nicht geeignet erachtet. Er wurde für diesen Zweck in internationaler Übereinkunft durch die begrifflich gänzlich anders zu verstehende Messunsicherheit abgelöst.

Die Ausführungen zur Messunsicherheit werden in Abschnitt 3.3 fortgesetzt.

1.4 Ziel der Arbeit

Das Konzept der Messunsicherheit bei der Datenauswertung soll in dieser Arbeit interdisziplinär auf die Psychologie übertragen werden. Es soll ein Verfahren entwickelt werden, das bei Anwendung der EDV praxisnah einer schnellen Erfassung und gewichteten Auswertung der beobachteten Daten eines aktuellen AC dienen kann. Dabei soll insbesondere auch für jeden einzelnen Teilnehmer die Unsicherheit zu den Ergebnissen ermittelt werden. Obwohl die Anwendungspraxis eines AC bei der Personalauswahl und Potenzialbeurteilung in den Vordergrund der Betrachtung gestellt wird, soll das Verfahren auch in ganz allgemeinen Urteils- und Entscheidungsprozessen Anwendung finden können. Daher ist im Folgenden der Begriff AC in einem sehr weiten, verallgemeinerten Sinne zu verstehen (Abschnitt 2.1).

Die Einführung und Quantifizierung der Messunsicherheit soll die folgenden Vorteile bieten:

- 1) Die Angabe der Messunsicherheit zur Bewertung jedes Merkmals eines einzelnen Teilnehmers oder zu dessen Gesamtbewertung in einem einzelnen durchgeführten AC drückt das Vertrauen in diese Bewertung oder dessen Genauigkeit oder Qualität auf der Basis der gewonnenen unvollständigen Information über den Teilnehmer aus. Sie quantifiziert den Mangel in dieser Information und damit auch das Risiko z.B. bei einer folgenden Personalentscheidung.

Dies wird auf folgende Weise erreicht: Erstens werden alle Merkmale wie physikalische Messgrößen behandelt und zweitens werden anders als üblich die Beobachter angewiesen, jeweils nicht nur eine Bewertung zu einem Merkmal eines Teilnehmers anzugeben, sondern deren zwei, nämlich eine minimale und eine maximale Bewertung, die nach Ansicht des Beobachters noch gerade ebenso gut dem beobachteten Verhalten des Teilnehmers angemessen sein kann. Die beiden Bewertungen dürfen gleich sein. Auf diese einfache Weise kann der Beobachter subjektiv jeweils seinem oft empfundenen Mangel an Information bei der Einschätzung der Ausprägung eines Merkmals bei einem Teilnehmer aus dessen beobachteten Verhalten heraus realistisch Ausdruck geben. So hat z.B. der Beobachter beim Bewerten in der Praxis häufig den Eindruck, dass er ein Merkmal nicht genau genug beobachten konnte oder sowohl positive als auch negative Aspekte bei diesem Merkmal feststellen konnte. Üblicherweise würde er in einem

solchen Fall eine mittlere Bewertung wählen, unabhängig von der Stärke seines subjektiven Eindrucks von Unsicherheit. Diesen Eindruck könnte er in der abschließenden Beobachterkonferenz nur noch verbal schildern. In der Bewertung selbst ist er nicht sichtbar. Die Angabe seiner Unsicherheit in Form einer für ihn noch als sinnvoll erachteten minimalen und maximalen Bewertung bietet dagegen die Möglichkeit, genau diesen Eindruck quantitativ auf einfache Weise festzuhalten und so als Daten in die Auswertung und gegebenenfalls in einen Entscheidungsprozess einfließen zu lassen. Die Angabe der beiden Bewertungen bedeutet somit ein Mehr an Information, das bei der Auswertung benutzt wird, um die Messunsicherheit zu berechnen. (Zu anderen Möglichkeiten der Bewertung siehe Abschnitt 4.3.3)

Stimmen beispielsweise die Gesamtbewertungen zweier Teilnehmer nahezu überein, so ist es nicht leicht, sich rational für einen von beiden zu entscheiden, ohne mit zusätzlichem Aufwand neue, zunächst gar nicht vorgesehene Entscheidungskriterien heranzuziehen. Wenn sich jedoch die Unsicherheiten zu den Bewertungen merklich unterscheiden, wird man denjenigen Teilnehmer auswählen dürfen, zu dessen Bewertung die geringere Unsicherheit gehört, und damit Risiko und Aufwand der Entscheidung möglichst gering halten können. Auch wenn sich die Gesamtbewertungen der Teilnehmer unterscheiden, kann man erst anhand der zugehörigen Unsicherheiten erkennen, ob der Unterschied signifikant ist (Weise und Wöger, 1994). Zwar macht es den Beobachtern mehr Mühe, jeweils Paare von Bewertungen anzugeben, und der Aufwand für die Auswertung ist höher. Das kann jedoch durch Einsatz der EDV kompensiert werden, vor allem, wenn die Beobachter alle ihre Bewertungen auf zweckmäßig gestalteten Formularblättern festhalten, die automatisch gelesen und ausgewertet werden können, oder auf geeignete Weise direkt in den Computer eingeben. In diesem Fall können alle berechneten Ergebnisse bereits bei der Beobachterkonferenz als Diskussionsgrundlage vorliegen.

Der genannte Vorteil wird besonders deutlich, wenn, wie es in der Praxis oft vorkommt, das AC singulär, also speziell für eine aktuelle wichtige oder dringliche komplexe Entscheidungsaufgabe konzipiert werden muss und nur einmal auf wenige Teilnehmer oder Alternativen anzuwenden ist. Dann besteht keine Möglichkeit, das AC mittels der konventionellen Statistik zu evaluieren. Umso wichtiger ist es dann, im AC möglichst viel Information zu sammeln und diese dann optimal auszuwerten. Beim physikalischen Messen hat sich dabei der Begriff der Messunsicherheit bewährt, der auf der Bayes'schen Statistik basiert. Diese benötigt nicht unbedingt statistische Information aus vielen Versuchen (Abschnitt 3.1).

- 2) Die Messunsicherheit liefert neue Validitätsmaße. Sie unterstützt die Untersuchung eines in Serie oder auf genügend viele Teilnehmer angewendeten AC-Verfahrens hinsichtlich der Konstruktvalidität, d.h. sie gibt durch Unsicherheitskennwerte Hinweise zu der Frage, ob eine angewendete Übung des AC überhaupt für die Bewertung eines bestimmten Merkmals geeignet ist. Sie ergänzt somit die üblichen Validitätsmaße, z.B. Korrelationskoeffizienten zur konvergenten und diskriminanten Konstruktvalidität. Außerdem können mittels der Unsicherheitskennwerte die einzelnen Merkmale hinsichtlich der Genauigkeit ihrer Ermittlung sowie das Bewertungsverhalten der Beobachter charakterisiert werden.

Wenn ein AC auf viele Teilnehmer angewendet worden ist oder auch viele Beobachter beteiligt worden sind, kann das AC durch Anwendung der konventionellen Statistik auf alle Ergebnisse wie üblich evaluiert werden (Kleinmann, 1997). Die Messunsicherheit kann die Evaluierung unterstützen: Wenn die Unsicherheit eines Merkmals in einer Übung, geeignet gemittelt über alle Teilnehmer und Beobachter, wesentlich kleiner oder größer ist als bei anderen Merkmalen, so wird man annehmen dürfen, dass die Übung für die Bewertung des Merkmals gut bzw. weniger gut geeignet ist. Wenn weiterhin die von einem Beobachter ausgedrückte Unsicherheit, geeignet gemittelt über alle Merkmale, Teilnehmer und Übungen, wesentlich kleiner oder größer ist als bei anderen Beobachtern, so kann dies Rückschlüsse auf die Urteilsfähigkeit jenes Beobachters erlauben. Die geeignet gemittelten Unsicherheiten sind die unter Punkt 2 genannten Unsicherheitskennwerte (Abschnitt 6.2).

Man könnte gegen die Unsicherheit eines Merkmals eines Teilnehmers einwenden, dass ja schon die Streuung der einzelnen Bewertungen zu diesem Merkmal ähnliche Aussagen ermögliche. Das ist richtig und dementsprechend enthält auch die zu errechnende Unsicherheit diese Streuung als eine ihrer Komponenten. Sie enthält aber noch weitere Komponenten, nämlich die vom Beobachter selbst beigetragene Unsicherheit aufgrund der jeweils angegebenen minimalen und maximalen Bewertungen, weiterhin die Unsicherheit, die durch die Stufung der Beurteilungsskala erzeugt wird, sowie die Unsicherheit, die den Gewichten zuzuordnen ist, mit denen Merkmale zu allgemeineren oder übergeordneten Merkmalen zusammengesetzt werden, z.B. zu einem Gesamtmerkmal, woran schließlich eine Entscheidung getroffen wird. Die neuen Validitätsmaße sind auch auf ein singuläres AC mit nur wenigen Teilnehmern anwendbar, sie sollten dann allerdings mit Vorsicht interpretiert werden.

1.5 Vorgehen

Im folgenden Kapitel 2 werden zunächst die Grundlagen des AC betrachtet, insbesondere die Grundbegriffe zum AC genau definiert, benannt und erläutert. Das ist nötig, weil das in dieser Arbeit zu entwickelnde Auswerteverfahren mit Berücksichtigung der Unsicherheit auch bei ganz allgemeinen Urteils- und Entscheidungsprozessen Anwendung finden soll. Für diesen Zweck sind die Begriffsinhalte zum AC, das exemplarisch im Vordergrund der Betrachtung steht, zu erweitern.

In Kapitel 3 werden danach die Grundlagen der Bayes'schen Statistik dargelegt, auf der diese Arbeit fußt. Die Messunsicherheit wird eingeführt und erläutert und es wird gezeigt, wie sie mit Hilfe von zwei wesentlichen Ansätzen auf das AC übertragen werden kann.

Das Kapitel 4 dient dem Aufbau des Verfahrens für die Auswertung der in einem AC gewonnenen Daten bis hin zur Ermittlung des Gesamtergebnisses für jeden Teilnehmer und der jeweils zugehörigen Unsicherheit. Es umfasst auch die Konstruktion eines AC, die Aufstellung des Modells der Auswertung, die Datenvorbereitung, die Interpretation der Unsicherheit sowie Vorschläge für die Implementierung des Verfahrens als Computer-Programmsystem.

Der Evaluierung dieses Auswerteverfahrens sind die Kapitel 5 und 6 gewidmet. In Kapitel 5 werden drei AC-Serien mit jeweils vielen Teilnehmern beschrieben, die für die Evaluierung herangezogen werden. Die eigentliche Evaluierung folgt in Kapitel 6 in drei Teilen.

In Abschnitt 6.1 werden die mit dem Auswerteverfahren dieser Arbeit gewonnenen Ergebnisse zu den Teilnehmern und deren Rangfolgen mit den Ergebnissen der üblichen Auswertung bzw. mit den Rangfolgen aus den Beobachterkonferenzen verglichen, wobei die Vorteile der zusätzlichen Angabe der Unsicherheit zu den Ergebnissen deutlich werden. Die Unsicherheitskennwerte, d.h. die neuen, auf der Unsicherheit beruhenden Validitätsmaße für die Evaluierung eines AC, werden in Abschnitt 6.2 eingeführt, berechnet und diskutiert. Schließlich werden in Abschnitt 6.3 Korrelationskoeffizienten betrachtet. Zum einen werden Korrelationskoeffizienten für die Untersuchung der Konstruktvalidität eines AC nach der konventionellen und nach der Bayes'schen Statistik berechnet und miteinander verglichen, zum anderen wird gezeigt, dass Korrelationen bei der Berechnung der Unsicherheiten vernachlässigbar sind.

Abschließend werden in Kapitel 7 die wesentlichen Aspekte und Ergebnisse dieser Arbeit zusammengefasst und es wird ein Ausblick auf mögliche sinnvolle weitere Untersuchungen gegeben.

Alle Ergebnisse der Evaluierung werden in Anhang A in Tabellen und Bildern dargestellt. In Anhang B wird ein vorläufiges Experimentier- und Demonstrationsprogramm für die Auswahl von Alternativen nach dem Auswerteverfahren für den praktischen Einsatz in einem aktuellen AC beschrieben.

2 Betrachtungen zu den Grundlagen des Assessment Centers

2.1 Allgemeines zur Verwendung von Begriffen

Aus zwei Gründen ist es erforderlich, die Grundlagen des AC genauer zu betrachten, insbesondere hinsichtlich Definition, Bedeutungsumfang und Benennung der Grundbegriffe.

Der erste Grund liegt darin, dass das in dieser Arbeit zu entwickelnde Auswerteverfahren unter Berücksichtigung der Messunsicherheit nicht nur auf psychologische AC im üblichen, engeren Sinne, sondern analog auch auf ganz allgemeine Urteils- und Entscheidungsprozesse im weitesten Sinne, auch auf anderen Fachgebieten, anwendbar sein soll. Das AC im üblichen Sinne spielt lediglich eine exemplarische Hauptrolle im Vordergrund der Betrachtungen. Damit dies nicht immer wieder herausgestellt werden muss, ist es für den Gebrauch in der vorliegende Arbeit nötig, die übliche Bedeutung der Grundbegriffe des AC zu verallgemeinern und den Bedeutungsumfang dieser Grundbegriffe stark zu erweitern. Es wird dabei jedoch wegen der exemplarischen Rolle des üblichen AC an einer AC-nahen Benennung der Grundbegriffe festgehalten. Dadurch erhält aber eine solche Benennung einen stellvertretenden, abstrakten Charakter. Das gilt schon für das *Assessment Center* selbst. Diese Benennung steht stellvertretend für jeden analogen Urteils- und Entscheidungsprozess. Dieser kann auf dem aktuellen Fachgebiet ganz anders genannt sein, etwas anderes bedeuten und auch anders ablaufen als ein übliches AC. Beispiel dafür ist ein Messverfahren in der Messtechnik. Wichtig ist nur, dass sich im aktuellen Fall eine Analogie zum üblichen AC auffinden lässt. Sie ist, wie dieses Beispiel zeigt, keineswegs immer offensichtlich. Gleiches gilt auch für die anderen Grundbegriffe. Sie stellen eigentlich Klassen von Begriffen dar, die umgangssprachlich oder fachspezifisch ganz unterschiedliche Bedeutung haben und auch ganz unterschiedlich benannt werden, die aber dennoch unter der verallgemeinernden Sicht dieser Arbeit als gleichartig oder analog zu betrachten sind. Die Begriffe Teilnehmer, Messobjekt und Alternative gehören z.B. zu einer solchen Klasse. Die Benennung *Teilnehmer* wird stellvertretend für alle Begriffe dieser Klasse benutzt.

Der zweite Grund ist, dass die Grundbegriffe des AC in der Literatur oft ohne genaue Definition sehr unterschiedlich benannt werden. So findet man z.B. für die Benennung *Merkmal* auch die Benennungen Konstrukt, Trait und Dimension, in mancher Publikation sogar wechselnd gebraucht, ohne dass klargestellt wird, was sie genau bedeuten

oder auch nur, dass sie dasselbe bedeuten, was den uneingeweihten Leser verwirren kann. Eine Verbesserung dahingehend bietet die neue Norm DIN 33430 (2002), in der informativ u.a. einige Grundbegriffe des AC erläutert werden.

Das sind die Gründe, warum im Folgenden im Interesse von Klarheit, begrifflicher Eindeutigkeit und interdisziplinärer Analogie und Verallgemeinerung die wichtigsten Grundbegriffe des AC genau definiert, AC-nah benannt, in ihrer Bedeutung erläutert sowie in ihrem Bedeutungsumfang umrissen werden. Die AC-nahen Benennungen werden durchgängig verwendet (auch schon in Kapitel 1), auch wenn das nicht immer passend ist, z.B. *Teilnehmer*, auch wenn es sich bei einer aktuellen Aufgabe gar nicht um eine Person handelt. Die Benennungen sind nötigenfalls durch die Benennungen der entsprechenden Begriffe auf dem aktuellen Fachgebiet zu ersetzen. Wenn das geschieht, müssen alle folgenden Definitionen der Grundbegriffe richtig bleiben. In diesem Sinne ist Analogie hier gemeint.

Neben den Grundbegriffen des AC werden in dieser Arbeit auch Begriffe der Wahrscheinlichkeitsrechnung und Statistik sowie der physikalischen Messtechnik verwendet. Im Gegensatz zu jenen sind diese Begriffe jedoch in Normen genau definiert, eindeutig benannt und festgelegt. Deshalb genügt es, die entsprechenden Normen zu zitieren. Die in dieser Arbeit benutzten Begriffe der Wahrscheinlichkeitsrechnung und Statistik sind in den Normen DIN 13303-1 und -2 (1982), DIN 55350-21 (1982) und DIN 55350-22 (1987) definiert, die Grundbegriffe der Messtechnik in den Normen DIN 1319-1 (1995), DIN 1319-3 (1996) und DIN 1319-4 (1999). Speziell zu den Definitionen der mit der Messunsicherheit verbundenen Begriffe siehe [Uns] und DIN 1319-1 (1995) sowie Abschnitt 3.3. Zur Bayes'schen Statistik siehe Abschnitt 3.1.

2.2 Definitionen und Erläuterungen zu den Grundbegriffen

2.2.1 Assessment Center

Der erste zu betrachtende Grundbegriff ist der des AC selbst. Er wird in dieser Arbeit so definiert:

- Ein *Assessment Center* (AC) besteht aus Übungen, in denen festgelegte Merkmale von Teilnehmern durch Beobachter bewertet werden.

Diese Definition des AC mutet zunächst fast trivial an. Sie gewinnt erst ihren umfassenden Sinn und lässt sich genauer und allgemein genug verstehen durch Definition und Interpretation der weiteren Grundbegriffe *Übung*, *Merkmal*, *Teilnehmer*, *Beobachter*

und *Bewertung*, worauf sich die Definition des AC stützt. Die genannten Grundbegriffe bilden ein Begriffssystem und werden in den folgenden Abschnitten dieses Kapitels 2 ausführlich behandelt. Jeder dieser Grundbegriffe und seine Benennung stehen, wie in Abschnitt 2.1 beschrieben, stellvertretend für eine ganze Klasse analoger Begriffe.

So steht die Benennung AC auch für ähnliche, analoge, aber ganz anders genannte Verfahren, wie Einstellungs-, Beurteilungs- und Messverfahren, Prüfung, Untersuchung, Entscheidungsanalyse. Dadurch gewinnt der Begriff *Assessment Center* eine sehr allgemeine Bedeutung.

Die Anzahl der Übungen, Merkmale, Teilnehmer und Beobachter darf jeweils gleich 1 sein, obwohl diese Begriffe in der Definition des AC im Plural stehen. Beispiel ist ein Interview unter vier Augen, das als AC mit nur einer einzigen Übung, einem Teilnehmer und einem Beobachter aufgefasst wird.

2.2.2 Teilnehmer

Die Definition des Teilnehmers lautet:

- Ein *Teilnehmer* ist eine Person oder ein Objekt, die bzw. das in einem AC hinsichtlich festgelegter Merkmale bewertet wird.

In einem AC im Personalbereich ist der Teilnehmer natürlich eine zu bewertende Person, z.B. ein Bewerber, Kandidat oder Konkurrent. In der physikalischen Messtechnik ist er ein Messobjekt, Träger der zu ermittelnden quantifizierbaren physikalischen Merkmale. Diese Merkmale sind die *Messgrößen*. Bei einer allgemeinen Entscheidungsaufgabe ist der Teilnehmer eine zu bewertende Alternative unter mehreren gleichartigen oder in Betracht gezogenen. Die Alternative kann sowohl eine Person als auch eine Mannschaft oder ein materielles oder ideelles Objekt sein, z.B. ein Gebäude, ein Prüfobjekt bzw. eine Vorgehensweise. Ganz allgemein betrachtet ist der Teilnehmer als Träger der Merkmale das Objekt der Bewertung im AC.

2.2.3 Beobachter

Ähnlich wie der Teilnehmer wird der Beobachter definiert:

- Ein *Beobachter* ist eine Person oder ein Objekt, die bzw. das die Teilnehmer in einem AC hinsichtlich festgelegter Merkmale bewertet.

In einem AC im Personalbereich ist der Beobachter meist eine bewertende Person, z.B. ein Gutachter, Prüfer, Juror, Wertungsrichter, Interviewer, Beisitzer. In der physikalischen Messtechnik entspricht dem Beobachter ein Messinstrument (Messgerät,

Messeinrichtung). Bei einer allgemeinen Entscheidungsaufgabe ist der Beobachter ein Entscheidungsträger. Manchmal wird ein Teilnehmer von einem Team gemeinsam bewertet, nicht von den Mitgliedern des Teams einzeln. In diesem Fall bildet das Team als Ganzes den Beobachter. Auch ein psychologischer Test, z.B. ein Intelligenztest, ist nicht nur als eine Übung nach Abschnitt 2.2.4 anzusehen, sondern gleichzeitig als Beobachter aufzufassen, wenn sein Testergebnis als Bewertung eines Merkmals benutzt wird. Falls Teilnehmer sich selbst oder gegenseitig bewerten (self reporting bzw. peer rating), was im AC vorgesehen sein kann, sind die Teilnehmer gleichzeitig auch Beobachter. Ganz allgemein betrachtet ist der Beobachter im AC das bewertende Subjekt. Wenn im Folgenden der Beobachter ausdrücklich als Person zu verstehen ist wie bei der Evaluierung in den Kapiteln 5 und 6, so wird dies besonders vermerkt oder er wird *Beobachterperson* genannt, um Missverständnisse zu vermeiden.

2.2.4 Übung

Die Übung wird als Vorgang wie folgt definiert:

- Eine *Übung* ist ein Vorgang, in dem die Teilnehmer in einem AC hinsichtlich festgelegter Merkmale bewertet werden.

Eine Übung ist sehr allgemein als ein Prozess zur Informationsgewinnung zu verstehen. Es kann sich dabei um eine beliebige Veranstaltung handeln, z.B. um ein Interview, einen Intelligenztest, ein Vorauswahlverfahren, einen Fachvortrag, einen Simulationstest am Computer oder eine gebräuchliche AC-Standardübung, wie eine Postkorb-Übung, eine Präsentation oder ein Rollenspiel. In der physikalischen Messtechnik ist die Übung eine Messung oder ein Experiment, d.h. die Durchführung eines geeigneten Messverfahrens. Bei einer allgemeinen Entscheidungsaufgabe ist die Übung die Durchführung eines Untersuchungs- oder Prüfverfahrens. In unterschiedlichen Übungen können dieselben oder unterschiedliche Merkmale bewertet werden.

2.2.5 Merkmal

Sehr wichtig ist der Grundbegriff Merkmal. Seine Definition lautet:

- Ein *Merkmal* ist eine im AC auf einer vereinbarten Bewertungsskala quantifizierte und zu bewertende Eigenschaft eines Teilnehmers.

Unter einem Merkmal im AC wird also immer ein quantitatives Merkmal verstanden. Der Teilnehmer ist der Träger dieses Merkmals. Wenn ein nichtquantitatives Merkmal zu bewerten ist, z.B. die Farbe, muss es quantifiziert werden. In jedem Fall müssen

alle möglichen Ausprägungen des Merkmals auf der Bewertungsskala so angeordnet werden, dass ein Skalenwert ausdrückt, wie wertvoll oder wichtig die zugehörige Ausprägung im Hinblick auf das Ziel des AC im Vergleich zu den anderen Ausprägungen ist. Es ist nicht unbedingt erforderlich, aber doch sehr zweckmäßig, in einem AC nur eine einzige Bewertungsskala für alle Merkmale zu verwenden. Das vereinfacht alle Berechnungen erheblich. Dementsprechend sind z.B. auch die möglichen Punktergebnisse eines Tests auf der Bewertungsskala geeignet anzuordnen. (Zur Konstruktion einer Bewertungsskala siehe Abschnitt 3.4.1, zur Wahl einer zweckmäßigen Skala siehe Abschnitt 4.3.2)

Andere in der Literatur gebräuchliche Benennungen für Merkmal im AC sind: Konstrukt, Trait, Dimension, Anforderungsdimension, Kompetenz, Auswahl- und Entscheidungskriterium. In dieser Arbeit wird die Benennung *Merkmal* benutzt, weil diese sehr allgemein ist und auch auf anderen Fachgebieten, z.B. in der Statistik, für eine zu untersuchende Eigenschaft eines Objekts üblich ist (zur Benennung Konstrukt siehe weiter unten). In der Physik ist ein quantitatives Merkmal eine physikalische Größe, speziell eine Messgröße, wenn sie sich auf ein Messobjekt bezieht und zu messen ist, z.B. die Temperatur des Badewassers kurz vor dem Bad.

Hier wird klar, dass zu bewertende Merkmale im AC wie physikalische Messgrößen aufgefasst werden. Dies ist einer der beiden wesentlichen Ansätze dieser Arbeit. Dieser Ansatz ermöglicht es bereits im Prinzip, den Begriff *Messunsicherheit* und alle dazu in der Physik erarbeiteten Konzepte, Methoden und Berechnungsverfahren direkt auf das AC zu übertragen.

Es wird vorausgesetzt, dass jeder Teilnehmer zumindest während der Durchführung des AC eine feste (wahre) Ausprägung eines im AC betrachteten Merkmals besitzt, die es zu ermitteln gilt. Das schließt nicht aus, dass sich die Ausprägung später ändern kann, z.B. Fachwissen durch Lernen. Bei Potenzialbeurteilungen ist es gerade das Ziel des AC, bei Merkmalen Stärken für besondere Anforderungen herauszufinden und Defizite zu erkennen, sodass diese z.B. durch anschließende Fortbildungsmaßnahmen vermindert werden können. Die Merkmale eines AC sind so zu konstruieren oder auszuwählen, dass die Voraussetzung als hinreichend erfüllt angesehen werden kann (Abschnitt 4.1).

Die Bewertungen eines Merkmals durch die Beobachter sind Schätzwerte für die jeweilige wahre Ausprägung. Es gibt Merkmale, die direkt von den Beobachtern bewertet werden, und solche, deren Bewertung indirekt aus den Bewertungen anderer Merkmale auf geeignete Weise, z.B. durch Mittelwertbildung, zu errechnen sind. Letztere sind zusammengesetzte Merkmale, u.U. mit mehreren anderen Merkmalen als Komponenten. Beispiel hierfür ist das Merkmal Intelligenz, wie es in manchen Intelligenztests

ermittelt wird. Auch operationalisierte Merkmale sind zusammengesetzte Merkmale, wenn die Komponenten dieser Operationalisierung, z.B. Verhaltensanker, selbst bewertet werden. Diese Komponenten können dann als eigene Merkmale angesehen und auch so genannt werden (Abschnitt 4.1).

Es ist jedoch nicht unproblematisch, auch für die Komponenten der Operationalisierung wie bei anderen Merkmalen jeweils eine feste Ausprägung während einer Übung vorauszusetzen. Beispielsweise können Ausprägungen sich zufällig oder durch Interaktionen von Teilnehmern untereinander oder mit anderen Übungspartnern situationsbedingt stark ändern. So kann die Komponente „Formuliert flüssig“ des operationalisierten Merkmals „Mündlicher Vortrag“ durch häufige Störungen beeinträchtigt werden. In solchen Fällen sollten aber wenigstens die übergeordneten operationalisierten Merkmale die Voraussetzung weitgehend erfüllen. Wegen dieser Problematik und auch, weil es die Berechnungen vereinfacht, ist es besser, wie die Erfahrungen bei der Durchführung der Evaluierung gezeigt haben, bewertete Komponenten nicht als eigene Merkmale, sondern ihre Bewertungen direkt als Bewertungen des zugehörigen operationalisierten Merkmals aufzufassen (Abschnitte 4.1, 4.3 und 6.1). Eine auf diese Weise behandelte und zu bewertende Komponente gleich welcher Art eines operationalisierten Merkmals wird im Folgenden *Verhaltensanker* genannt, auch wenn die Komponente sich gar nicht auf das Verhalten einer Person bezieht, sondern z.B. nur einen Teilaspekt des Merkmals darstellt. In unterschiedlichen Übungen können zu demselben operationalisierten Merkmal durchaus unterschiedliche Verhaltensanker zu bewerten sein, damit alle interessierenden Verhaltensweisen oder Teilaspekte des Merkmals abgedeckt werden, wenn dies nicht durch eine einzige Übung möglich ist (Abschnitt 4.1).

Zur Verwendung der Benennung Konstrukt ist eine Bemerkung an dieser Stelle angebracht. Unter einem Konstrukt wird ein nicht direkt beobachtbares Merkmal (latente Variable) verstanden, auf das aus anderen Konstrukten und direkt beobachtbaren Merkmalen auf geeignete Weise mittels eines Modells (nomologisches Netzwerk) indirekt zu schließen ist (Kleinmann, 1997). Jedes zusammengesetzte Merkmal, wie oben erwähnt, ist demnach ein Konstrukt. Dies gilt auch für jedes operationalisierte Merkmal, weil darauf aus den beobachteten Komponenten geschlossen wird. Da nun im AC jedes Merkmal, wenn sinnvoll, operationalisiert werden sollte, besteht zwischen Merkmal und Konstrukt nur ein geringfügiger begrifflicher Unterschied, der im Folgenden vernachlässigt wird. Es wird deshalb nicht mehr von Konstrukten gesprochen. Allerdings wird der eingeführte Ausdruck *Konstruktvalidität* nicht durch Merkmalsvalidität ersetzt.

Für die Bewertung und wahrscheinlichkeitsmäßige Untersuchung wird jedem Merkmal bezüglich jedes einzelnen Teilnehmers und über alle Übungen hinweg, in denen das Merkmal zu bewerten ist, eine Zufallsvariable zugeordnet, ein *Schätzer*. Die Teilnehmer besitzen also zu demselben Merkmal je einen eigenen Schätzer. Jeder gewonnene Wert des Schätzers, z.B. jede Bewertung des Merkmals oder eines zugehörigen Verhaltensankers, ist ein *Schätzwert* für die zu ermittelnde wahre Ausprägung des Merkmals beim Teilnehmer. Ein Schätzer und der Einfachheit halber auch das zugehörige Merkmal werden im Folgenden mit demselben großen Buchstaben X, Y oder Z bezeichnet, ein Schätzwert dazu mit dem entsprechenden kleinen Buchstaben x, y bzw. z.

Manchmal hängt die Zuordnung von Schätzern zu einem Merkmal eines Teilnehmers von der Fragestellung ab. Wenn mehrere Bewertungen eines Merkmals in einer Sequenz über eine längere Zeitspanne vorliegen und angenommen werden darf, dass sich die wahre Ausprägung des Merkmals beim Teilnehmer in dieser Zeitspanne nicht geändert hat, genügt es, dem Merkmal einen einzigen Schätzer zuzuweisen und alle Bewertungen als Werte dieses Schätzers anzusehen. Die Reihenfolge der Bewertungen ist dann unerheblich. Wenn jedoch die zeitliche Änderung der wahren Ausprägung des Merkmals beim Teilnehmer untersucht werden soll (z.B. bei Bewertungen on the job), muss dem Merkmal zu jedem Bewertungszeitpunkt ein eigener Schätzer zugewiesen werden. Es ist dann so, als gehörten zu den Bewertungszeitpunkten unterschiedliche Merkmale von gleicher Art.

Während es in einem laufenden AC darum geht, die wahre Ausprägung mehrerer Merkmale bei jedem einzelnen Teilnehmer zu ermitteln, interessieren bei der Evaluierung über alle Teilnehmer oder auf andere Weise gemittelte Größen. Deshalb ist es bei der Evaluierung erforderlich, Zufallsvariablen der jeweiligen Fragestellung entsprechend einzuführen. Um z.B. für die Untersuchung der Konstruktvalidität Korrelationskoeffizienten $\varrho(X, Y)$ berechnen zu können, muss einem betrachteten Merkmal für a l l e Teilnehmer in einer bestimmten Übung dieselbe Zufallsvariable X zugewiesen werden und demselben Merkmal für alle Teilnehmer in einer anderen Übung ebenso dieselbe, aber von X verschiedene Zufallsvariable Y. Alle Bewertungen des Merkmals für alle Teilnehmer in den beiden Übungen sind dann Realisierungen der Zufallsvariablen X bzw. Y. (Abschnitt 6.3)

2.2.6 Bewertung

Schließlich ist noch die Bewertung als Grundbegriff zu definieren:

- Als *Bewertung* wird ein Schätzwert aufgefasst, der nach einer Übung im AC von einem Beobachter auf einer vereinbarten Bewertungsskala für die zu ermittelnde (wahre) Ausprägung eines Merkmals bei einem Teilnehmer abgegeben wird.

Die Bewertung wird oft z.B. auch Benotung, Beurteilung, Rating, Schulnote, Votum, Zensur genannt, beim physikalischen Messen ist sie ein Messwert oder ein Messergebnis für eine Messgröße. Sie stellt ganz allgemein im AC gewonnene Information dar.

Das in dieser Arbeit vorzuschlagende Verfahren für die Auswertung der im AC gewonnenen Information unterscheidet sich wesentlich in der Art der Bewertung von anderen Verfahren. Die Beobachter werden angewiesen, statt jeder einzelnen Bewertung für ein Merkmal eines Teilnehmers jeweils eine minimale und eine maximale Bewertung abzugeben, die auch beide gleich sein dürfen. Dadurch wird ein Bereich derjenigen möglichen Ausprägungen des Merkmals abgegrenzt, die nach Ansicht und Dafürhalten des Beobachters aufgrund der in der Übung gewonnenen Information über den Teilnehmer vernünftigerweise im gleichen Maße als realistische Schätzwerte für das Merkmal infrage kommen. Die Angabe jeweils einer minimalen und einer maximalen Bewertung anstelle einer einzelnen Bewertung bedeutet ein Mehr an Information. Diese wird dazu benutzt, nicht nur den besten Schätzwert für das Merkmal zu bilden, sondern auch die Unsicherheit des Beobachters bei der Abgabe seiner Bewertung wegen mangelnder Information aus der Übung oder aus eigenen Gründen des Beobachters zu quantifizieren. Der Beobachter darf die Bewertung auch verweigern, wenn er meint, dass die erhaltene Information in keiner Weise für eine vernünftige Bewertung ausreicht. In diesem Fall kommen alle Skalenwerte als Schätzwerte für die Ausprägung des gerade zu bewertenden Merkmals beim Teilnehmer gleichermaßen in Betracht. Die Angabe einer minimalen und einer maximalen Bewertung als zusammengehöriges Paar durch einen Beobachter im AC wird kurz *Aussage* des Beobachters genannt.

Die eben beschriebene Art der Bewertung bildet den zweiten wesentlichen Ansatz dieser Arbeit. Er dient dazu, im AC die für die Quantifizierung der Messunsicherheit nötige Information zu gewinnen.

Die Ausführungen zu Bewertungen werden in den Abschnitten 4.3.3 und 4.3.4 fortgesetzt.

3 Bayes-statistische Grundlagen

3.1 Bayes'sche und konventionelle Statistik, Wahrscheinlichkeit

Die *Bayes'sche Statistik* (z.B. Lee, 1989; Wickmann, 1990) ist Grundlage dieser Arbeit. Es wird deshalb kurz beschrieben, was an ihr und an der üblicherweise angewendeten *konventionellen Statistik* charakteristisch ist und worin sie sich im Wesentlichen unterscheiden.

Konventionelle und Bayes'sche Statistik fußen beide auf der Wahrscheinlichkeitsrechnung. Sie unterscheiden sich jedoch darin, wie *Wahrscheinlichkeit* zu verstehen ist. Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A wird in der konventionellen Statistik als *relative Häufigkeit* aufgefasst, mit der das Ereignis in vielmals unabhängig wiederholten Versuchen auftritt. Dagegen wird sie in der Bayes'schen Statistik als relative Anzahl der alternativen gleichartigen Möglichkeiten interpretiert, mit denen zusammen das Ereignis in einem Versuch eintreten kann, b e v o r dieser Versuch überhaupt durchgeführt oder sein Ergebnis zur Kenntnis genommen wird. Dies ist die *klassische Wahrscheinlichkeit* nach Bernoulli und Laplace (Laplace, 1820). Dieser Begriff entstand aus Untersuchungen über Glücksspiele.

Die erwähnten alternativen gleichartigen Möglichkeiten, von denen aufgrund vorliegender Information in jedem Versuch genau eine eintreten kann, sind die *Elementarereignisse*. Jedem von diesen wird in der Bayes'schen Statistik aus Prinzip die gleiche Wahrscheinlichkeit zugewiesen. Für dieses Prinzip sind mehrere Namen gebräuchlich, u.a. *Bernoulli'sches Prinzip* oder *Prinzip der gleichen Apriori-Wahrscheinlichkeiten*, in verallgemeinerter Form auch *Prinzip der maximalen (Informations-)Entropie* (PME). Danach erhält man die Wahrscheinlichkeit $1/6$ für die Augenzahl jeder Seite eines Würfels allein aus der Vorstellung des Würfels und der Art und Weise, den Würfel zu werfen, und ohne dass der Würfel wirklich geworfen wird. Oder man erhält die Wahrscheinlichkeit $1/n$ für jedes von n gleichartigen Losen in einer Lostrommel, bevor ein Los gezogen und dessen Inhalt zur Kenntnis genommen wird. Die klassische Wahrscheinlichkeit $P(A)$ kann dementsprechend als Chance zum Wetten auf das Eintreten eines Ereignisses A verstanden werden.

Während die konventionelle Statistik für Wahrscheinlichkeitsaussagen im Wesentlichen nur *statistische Information* aus wiederholt durchgeführten Versuchen verwendet, benutzt die Bayes'sche Statistik dafür auch auf andere Weise gewonnene *nichtstatistische*

Information. Die Bayes'sche Statistik wird manchmal *subjektive Statistik* genannt, weil neben gegebener Information, zu der auch die gewonnenen Ergebnisse aus wiederholt durchgeführten Versuchen gehören können, auch oft subjektive Einschätzungen einfließen. Diese sollten aber immer auf vorliegenden Fakten, begründeten Annahmen und Erfahrung beruhen. Die konventionelle Statistik wird auch *Häufigkeitsstatistik* oder *objektive Statistik* genannt, obwohl auch bei ihrer Anwendung oft Wahrscheinlichkeitsverteilungen, d.h. Häufigkeitsverteilungen eintretender Ereignisse, z.B. Normalverteilungen, angenommen oder angesetzt werden. Deren Parameter, z.B. Erwartungswerte, sind dann zu schätzen. Wahrscheinlichkeitsverteilungen in der Bayes'schen Statistik dürfen dagegen im Allgemeinen n i c h t als Häufigkeitsverteilungen aufgefasst werden, sondern stellen den auf der gerade vorliegenden Information beruhenden Kenntnisstand (degree of belief) bezüglich des Eintretens der möglichen Ereignisse dar. Sie können nach obigem Prinzip aufgestellt werden. Die Parameter dieser Verteilungen sind daher b e k a n n t.

Die konventionelle Statistik lässt sich gut anwenden, wenn hinsichtlich einer zu untersuchenden Frage ein genügend umfangreiches Datenmaterial aus vielmals wiederholten Versuchen verfügbar ist. Beide Statistiken kommen in diesem Fall, wenn analog vorgegangen wird und die zugrunde gelegten Annahmen zutreffen, sogar asymptotisch, d.h. bei unbeschränkt vielen Versuchen, zu denselben Ergebnissen (Abschnitt 3.4.2). Die Bayes'sche Statistik hat aber den Vorteil, dass sie auch im Fall geringer Information, z.B. unzureichender Daten, und selbst bei unsicheren Einflüssen, die sich bei wiederholten Versuchen nicht zufällig ändern, Wahrscheinlichkeitsaussagen ermöglicht. Gerade in der Praxis der Personalauswahl und Potenzialbeurteilung sowie bei allgemeinen Urteils- und Entscheidungsprozessen ist der Fall konventionell-statistisch unzureichender Daten die Regel. Meist liegt hinsichtlich eines wichtigen Merkmals eines Teilnehmers in einem AC nur die unsichere Bewertung eines oder weniger Beobachter vor. Die konventionelle Statistik ist in einem solchen Fall nicht anwendbar.

Diese Tatsache hat im Bereich der Personalauswahl und Potenzialbeurteilung dazu geführt, dass oft nicht quantitative, sondern beschreibende Bewertungsprofile angegeben werden, auch mit dem Argument, dass genaue quantitative Aussagen nicht möglich seien, z.B. beim Einstellungsinterview. Die Bayes'sche Statistik erlaubt nun Wahrscheinlichkeitsaussagen zu einem Bewertungsprofil und auch die Quantifizierung der Unsicherheit. Sie steht mit ihrer Betrachtungsweise daher zwischen der rein quantitativen und der rein beschreibenden Art der Bewertung. Wegen der Gründung der Bayes'schen Statistik auf der Wahrscheinlichkeitsrechnung lässt sich zur Unterstützung der Bewertung bei der Datenauswertung die EDV anwenden.

Wenn auch die Bayes'sche Statistik für die Betrachtungen dieser Arbeit besser geeignet erscheint als die konventionelle Statistik und deshalb hier zugrunde gelegt wird, soll das keine Abwertung der konventionellen Statistik bedeuten. In Fällen, in denen beide Statistiken anwendbar sind, ergänzen sie sich. Dann folgt aus ihrer asymptotischen Äquivalenz bei analogem Vorgehen und zutreffenden Annahmen, dass sich theoretische Annahmen der Bayes'schen Statistik mittels der konventionellen Statistik experimentell überprüfen lassen, sofern ausreichende Information dafür aus genügend oft wiederholten Versuchen gewonnen werden kann. So lässt sich die Wahrscheinlichkeit $1/6$ in der Bayes'schen Statistik für die Augenzahl jeder Seite eines zunächst als ideal angenommen Würfels durch vielmaliges Würfeln überprüfen. Wenn dabei die relative Häufigkeit des Auftretens einer betrachteten Augenzahl signifikant von $1/6$ abweicht, kann daraus folgen, dass der Würfel gezinkt ist.

3.2 Gedankenexperimente zur Informationsgewinnung

Einige einfache Gedankenexperimente zur Informationsgewinnung sollen an die Bayesstatistischen Grundlagen heranführen. Diese Gedankenexperimente betreffen geworfene, aber unaufgedeckte Würfel, wenn sich Information über die geworfenen Augenzahlen nur unvollständig ermitteln lässt. Es wird erläutert, wie unterschiedliche, womöglich widersprüchliche Aussagen von Beobachtern über die geworfenen Augenzahlen vernünftig und widerspruchsfrei zusammenzufassen sind. Dazu wird die (klassische) Wahrscheinlichkeitsverteilung für die Augenzahlen aufgestellt. Schließlich werden die Augenzahlen mit Merkmalen identifiziert.

3.2.1 Zusammenfassung von Information

Ein Würfel werde mit einem Würfelbecher geworfen, dieser aber noch nicht abgehoben. Ein Wetter verlässt sich auf einen Beobachter, der den Becher zwar leicht anheben kann, aber vielleicht wegen schlechter Beleuchtung nur zu der Aussage A kommt: „Ich würde nur auf 3, 4 oder 5 setzen“. Daraufhin weist der Wetter jeder dieser drei in der Aussage genannten alternativen gleichartigen Möglichkeiten nach dem Bernoulli'schen Prinzip aus Abschnitt 3.1 die gleiche Wahrscheinlichkeit $1/3$ zu. Ohne den Würfel erneut zu werfen, komme ein zweiter Beobachter, der nichts von der Aussage des ersten weiß, auf gleiche oder auch andere Weise unabhängig zu der Aussage B: „Ich würde nur auf 5 oder 6 setzen“. Aufgrund allein dieser Aussage setzt der Wetter für jede der beiden darin genannten alternativen Möglichkeiten die Wahrscheinlichkeit $1/2$ an. Aber welche Wahrscheinlichkeiten wird der Wetter den Augenzahlen zuordnen, wenn er die Aussagen beider Beobachter zusammen berücksichtigt? Diese Frage, wie Information zusammenzufassen ist, ist von grundlegender Bedeutung für diese Arbeit.

Der Wetter ist ein Spieler, er könnte deshalb eine der beiden Aussagen A und B durch Werfen einer Münze auswählen. Er ordnet auf diese Weise jeder der beiden Aussagen die Wahrscheinlichkeit $P(A) = P(B) = 1/2$ für die Auswahl zu. Jetzt lässt sich der Entwicklungssatz der Wahrscheinlichkeitsrechnung anwenden, um die Wahrscheinlichkeit $P(X)$ für eine mögliche Augenzahl X des Würfels zu finden:

$$P(X) = P(X|A) \cdot P(A) + P(X|B) \cdot P(B) + \ldots \tag{1}$$

Die Pünktchen ... bedeuten, dass entsprechende Glieder hinzukommen, wenn Aussagen C, D, ... weiterer Beobachter vorliegen. Dann ist bei n Aussagen natürlich $P(A) = P(B) = P(C) = \ldots = 1/n$ anzusetzen. In Gleichung (1) sind $P(X|A)$ und $P(X|B)$ usw. die schon oben angesetzten (bedingten) Wahrscheinlichkeiten der Augenzahlen bezüglich der Aussagen A bzw. B usw. allein. Diese Aussagen schließen sich durch das Werfen der Münze oder z.B. durch Ziehen eines Loses aus einer Lostrommel, in der für jede der n Aussagen ein Los liegt, gegenseitig aus. Das ist wichtig, weil Gleichung (1) nur dann gilt. Daraus ergeben sich nun für die Augenzahlen $X = 1$ bis 6 die Wahrscheinlichkeiten

$$\begin{aligned} P(1) &= P(2) = 0 \\ P(3) &= P(4) = (1/3) \cdot (1/2) + 0 \cdot (1/2) = 1/6 \\ P(5) &= (1/3) \cdot (1/2) + (1/2) \cdot (1/2) = 5/12 \\ P(6) &= 0 \cdot (1/2) + (1/2) \cdot (1/2) = 1/4 \end{aligned} \tag{2}$$

Man könnte einwenden, dass die beiden Aussagen A und B zusammen ja nur den exakten Schluss auf die Augenzahl $X = 5$ zulassen. Diese Schlussweise bleibt aber dann ergebnislos, versagt also, wenn, was vorkommen kann, die Aussagen sich widersprechen, B z.B. lauten würde: „Ich würde nur auf 1 oder 2 setzen“. Es wird betont, dass die Aussagen n i c h t bedeuten, dass die Augenzahl wirklich im Fall von Aussage A gleich 3, 4 oder 5 und im Fall von Aussage B gleich 5 oder 6 ist. Die Aussagen „Ich würde nur auf 5 oder 6 setzen“ und „Die Augenzahl beträgt 5 oder 6“ sind also nicht identisch und deshalb sorgsam zu unterscheiden. Die Aussagen „Die Augenzahl beträgt 3, 4 oder 5“ und „Die Augenzahl beträgt 5 oder 6“ würden tatsächlich nur den Schluss auf die Augenzahl 5 zulassen (Schnittmenge der Mengen $\{3, 4, 5\}$ und $\{5, 6\}$). Die Wahrscheinlichkeiten sind aber nur als Einschätzungen aufgrund der vorhandenen Information für eine vernünftige Wette zu verstehen. Im betrachteten Fall bedeutet dies nach Gleichung (2) eine größere Wahrscheinlichkeit der Augenzahl 5 im Vergleich zu den anderen Wahrscheinlichkeiten, weil nur die Augenzahl 5 in beiden Aussagen vorkommt. Ein Resultat, das vernünftig erscheint.

Eine andere Möglichkeit könnte darin bestehen, die Aussagen A und B der beiden Beobachter zur Gesamtaussage „Ich würde nur auf 3 bis 6 setzen“ zusammenzufassen (Vereinigung der Mengen $\{3, 4, 5\}$ und $\{5, 6\}$) und $P(X) = 1/4$ für jede dieser Augenzahlen $X = 3$ bis 6 anzusetzen. Aber auf diese Weise würde sich $P(5)$ nicht vergrößern, was man dann erwarten sollte, wenn mehrere weitere Beobachter jeweils dieselbe Aussage C „Ich würde nur auf 5 setzen“ hinzufügen würden, denn die Gesamtaussage „Ich würde nur auf 3 bis 6 setzen“, die alle genannten Augenzahlen zusammenfasst, ändert sich dadurch nicht. Nach Gleichung (1) dagegen vergrößert sich $P(5)$ erwartungsgemäß wie folgt, wenn n Aussagen C hinzukommen: Insgesamt liegen dann $m = n + 2$ Aussagen vor, sodass

$$P(5) = \frac{1}{3} \cdot \frac{1}{m} + \frac{1}{2} \cdot \frac{1}{m} + n \cdot \frac{1}{1} \cdot \frac{1}{m} = \frac{n + 5/6}{n + 2} > \frac{5}{12} \tag{3}$$

Bei größer werdender Anzahl n der hinzugefügten Aussagen C konzentriert sich die Wahrscheinlichkeit immer mehr auf die Augenzahl 5, deren Wahrscheinlichkeit nähert sich dem maximalen Wert 1. Es wird dann fast sicher, dass die Augenzahl 5 tatsächlich vorliegt. Ob das aber wirklich so ist, bleibt verborgen, solange der Würfelbecher nicht abgehoben wird.

Eine weitere Möglichkeit könnte man darin sehen, Wahrscheinlichkeiten dafür, dass die Aussagen A und B richtig oder falsch sind, zu berücksichtigen. Für die Festlegung dieser Wahrscheinlichkeiten liegt jedoch keinerlei Information vor. Man könnte aber trotzdem versuchen, sie den Möglichkeiten entsprechend festzulegen: z.B., wenn nur B betrachtet wird, $P(B) = 1/3$ dafür, dass B richtig ist, und $P(\overline{B}) = 1 - P(B) = 2/3$ dafür, dass B falsch ist, weil B nur 2 von den 6 möglichen Augenzahlen nennt. $\overline{B}$ ist die entgegengesetzte Aussage zu B, nämlich „Ich würde nicht auf 5 oder 6 setzen“. Sie entspricht der Aussage „Ich würde nur auf 1 bis 4 setzen“. Für jede Augenzahl X ergibt sich aber in diesem Fall analog Gleichung (1) $P(X) = P(X|B)P(B) + P(X|\overline{B})P(\overline{B}) = 1/6$, was man leicht nachprüfen kann. Das ist gleichbedeutend damit, dass gar keine Information gegeben ist, also die Aussage B gar nicht genannt ist. Denn wenn gar keine Aussagen vorliegen, also keinerlei Information gegeben ist, gilt immer für jede Augenzahl X die Wahrscheinlichkeit $P(X) = 1/6$. Diese Überlegung zeigt auch, dass sich Aussagen in ihrem Informationsgehalt gegenseitig aufheben können, hier B und $\overline{B}$. Auf gleiche Weise heben sich auch die sechs unterschiedlichen Aussagen „Ich würde nur auf i setzen“ mit $i = 1$ bis 6, wenn sie alle vorliegen, in ihrem Informationsgehalt gegenseitig auf.

Nach den Ausführungen dieses Abschnitts erweist sich unter den betrachteten Möglichkeiten der Zusammenfassung von Information nur die nach Gleichung (1) als sinnvoll.

Diese lässt auch widersprüchliche Aussagen A und B zu, weil sie nur Wahrscheinlichkeiten der Augenzahl betrifft. Die Aussagen sind aber immer als Einschätzung von Chancen wie beim Wetten zu formulieren und zu interpretieren.

3.2.2 Unabhängigkeit

Der Versuch in Abschnitt 3.2.1 wird nun mit zwei Würfeln durchgeführt und es wird nach den Augenzahlen X und Y der beiden Würfel gefragt. Der Beobachter gibt Aussagen A und B zu beiden Augenzahlen ab, z.B. A „Ich würde bei X nur auf 3 bis 5 setzen“ und B „Ich würde bei Y nur auf 5 oder 6 setzen“. Diese Aussagen können auch von zwei unterschiedlichen Beobachtern stammen. Der Wetter setzt nun wieder nach dem Bernoulli'schen Prinzip für gleichartige alternative Möglichkeiten gleiche Wahrscheinlichkeiten an, d.h. für die $3{\cdot}2 = 6$ infrage kommenden Paare (3,5), (3,6), (4,5), (4,6), (5,5), (5,6) der Augenzahlen (X, Y) dieselbe Wahrscheinlichkeit $P(X, Y|A, B) = 1/6$ und für die übrigen 30 Paare, die nicht infrage kommen, die Wahrscheinlichkeit null. Unabhängig davon setzt der Wetter auch die Wahrscheinlichkeiten $P(X|A) = 1/3$ und $P(Y|B) = 1/2$ wie im vorangehenden Abschnitt an. Nun erweist es sich aber nicht nur in diesem Beispiel für A und B, sondern ganz allgemein, wie man leicht nachrechnen kann, dass

$$P(X, Y|A, B) = P(X|A) \cdot P(Y|B) \tag{4}$$

Das bedeutet aber, dass die Augenzahlen X und Y u n a b h ä n g i g und damit auch u n k o r r e l i e r t sind.

Die Unabhängigkeit der Augenzahlen der beiden Würfel folgt allein aus der Anwendung des Bernoulli'schen Prinzips. Die Würfel könnten zwar wie Magnete dazu neigen, in bestimmter Lage aneinander zu haften, oder der Beobachter könnte eine Vorliebe für die Zahl 5 haben, aber das ist nicht bekannt und bleibt auch verborgen. Darüber liegt keinerlei Information vor und kann deshalb in der Bayes'schen Statistik auch nicht benutzt werden, weil diese sich n u r auf vorhandene Information und das Prinzip der Zuordnung von Wahrscheinlichkeiten stützt. Manch einen mag dies verwundern, der im Rahmen der konventionellen Statistik zu denken gewohnt ist. Man muss aber bedenken, dass hier nicht vielmals wiederholt gewürfelt wird und der Würfelbecher auch nicht abgehoben wird, wobei ein Haften der Würfel aneinander erst offenbar werden würde. Als Information liegen dem Wetter nur die Aussagen A und B bezüglich der Augenzahlen X bzw. Y der beiden Würfel vor und sonst gar nichts. Hier tritt zwar keine Korrelation auf, das bedeutet aber nicht, dass das allgemein so gilt.

Beim obigen Versuch mit zwei Würfeln wird nun nach der Summe $Z = X + Y$ der Augenzahlen X und Y der beiden Würfel gefragt. Mehrere Beobachter geben

jeweils Aussagen zu beiden Augenzahlen ab und zwar unabhängig voneinander und jeder Beobachter unabhängig bezüglich der beiden Würfel entsprechend Gleichung (4). Letzteres folgt wieder, weil zur Beurteilung der Abhängigkeit der beiden Aussagen eines Beobachters voneinander bezüglich der beiden Würfel keinerlei Information vorliegt. Der Wetter ermittelt sowohl $P(X)$ als auch $P(Y)$ nach Gleichung (1).

Wie lautet nun die gemeinsame Wahrscheinlichkeit $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$? Zur Beantwortung dieser Frage seien für einen Moment alle Aussagen der Beobachter über Y vergessen oder beliebig verändert. Dadurch ändert sich an der Information über X gar nichts und es folgt, dass die bedingte Wahrscheinlichkeit $P(X|Y)$ nicht davon abhängt, welchen Wert Y gerade hat, d.h. es ist $P(X|Y) = P(X)$, woraus

$$P(X,Y) = P(X) \cdot P(Y) \tag{5}$$

folgt. Die Augenzahlen X und Y sind also unabhängige und damit auch unkorrelierte Zufallsvariablen. Es ist also die Kovarianz $\mathrm{Cov}\,(X,Y) = 0$. Zu betonen ist, dass die Unabhängigkeit von X und Y allein aus der besonderen Form der Aussagen A und B folgt. Eine Aussage wie „Ich würde nur bei X auf 3 oder 4 und bei Y auf 5 setzen oder aber bei X auf 5 und in diesem Fall bei Y auf 5 oder 6“ würde die Unabhängigkeit zerstören. Denn hier hängt die Aussage zu X davon ab, welcher Wert Y zugewiesen wird. Setzt man auf $Y = 5$, so lautet die Aussage zu X „Ich würde bei X nur auf 3, 4 oder 5 setzen“, setzt man dagegen auf $Y = 6$, so ändert sich die Aussage zu X in „Ich würde bei X nur auf 5 setzen“.

Auch die Summe $Z = X + Y$ ist eine Zufallsvariable. Ihre Wahrscheinlichkeit $P(Z)$ lässt sich zwar ebenfalls berechnen, jedoch wird sie nicht benötigt werden. Wichtig sind aber die folgenden Beziehungen zwischen den Erwartungswerten (E), den Varianzen (Var) und der Kovarianz (Cov) der Zufallsvariablen:

$$\mathrm{E}\,Z = \mathrm{E}\,X + \mathrm{E}\,Y \tag{6}$$

$$\mathrm{Var}\,(Z) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) + 2\,\mathrm{Cov}\,(X,Y) \tag{7}$$

Sind X und Y unkorreliert, gilt $\mathrm{Cov}\,(X,Y) = 0$. Das ist der Fall, wenn X und Y unabhängige Zufallsvariablen sind.

3.2.3 Übertragung auf Merkmale

Für die Anwendung im AC werden die Gedankenexperimente in den Abschnitten 3.2.1 und 3.2.2 einfach uminterpretiert: Der Würfelbecher mit einem geworfenen, aber verborgenen Würfel wird nun angesehen als ein Teilnehmer mit einem zu bewertenden

Merkmal, z.B. der Kommunikationsfähigkeit. Dieses Merkmal kann beim Teilnehmer mehr oder weniger ausgeprägt sein, was der unbekannten geworfenen Augenzahl des Würfels entspricht. Die Ausprägung des Merkmals sei durch die beiden Beobachter nach einer Übung im AC unabhängig voneinander auf der gestuften Skala von 1 bis 6 mit der Stufenhöhe 1 zu bewerten. Die Übung kann immer nur ein unvollkommener Test für das zu bewertende Merkmal sein, die wirkliche (wahre) Ausprägung dieses Merkmals beim Teilnehmer bleibt verborgen – hier kann der „Würfelbecher“ nicht abgehoben werden. Deshalb sollen die Beobachter aufgrund des in der Übung tatsächlich gezeigten Verhaltens des Teilnehmers nur die Aussagen A bzw. B wie in Abschnitt 3.2.1 als ihre Bewertungen abgeben können. Die gewonnene, unvollständige Information über das Merkmal wird also durch die Bewertungen der Beobachter gebildet. Die Bewertungen können sich widersprechen.

Der Fall, dass zwei Merkmale zu bewerten sind, entspricht auf gleiche Weise dem Fall zweier Würfel in Abschnitt 3.2.2. So lassen sich die Abschnitte 3.2.1 und 3.2.2 unmittelbar auf das hier betrachtete einfache, aber grundlegende Beispiel eines AC mit zwei Merkmalen übertragen, insbesondere die Wahrscheinlichkeitsaussagen über die Augenzahlen der Würfel auf die zu bewertenden Merkmale.

In einem weiteren Schritt wird die Summe Z der beiden Merkmale dieses Beispiels analog Abschnitt 3.2.2 als zusammengesetztes Merkmal interpretiert, z.B. als die Gesamtkompetenz des Teilnehmers, die bei der Personalauswahl mit den Gesamtkompetenzen anderer Teilnehmer zu vergleichen ist. Z kann ganzzahlige Werte von 2 bis 12 annehmen. Aber wie können die Gesamtkompetenzen der Teilnehmer auf einfache Weise kritisch miteinander verglichen werden, wenn zunächst jeweils nur Wahrscheinlichkeitsaussagen über die möglichen Ausprägungen der beiden Merkmale bei jedem einzelnen Teilnehmer vorliegen? Diese Frage wird im folgenden Abschnitt 3.3 aufgegriffen und beantwortet.

3.3 Einführung der Messunsicherheit

Mit Hilfe des Begriffs der Messunsicherheit, der schon in Abschnitt 1.1 erwähnten interdisziplinären Anleihe aus der physikalischen Messtechnik, soll die am Ende des Abschnitts 3.2.3 gestellte Frage nach dem kritischen Vergleich von Merkmalen in Abschnitt 3.3.3 Beantwortung finden. Die Messunsicherheit wird in Abschnitt 3.3.1 qualitativ definiert und in Abschnitt 3.3.2 durch Einführung der Standardunsicherheit quantifiziert. Siehe hierzu auch Abschnitt 1.3 und [Uns] sowie eine kurze Zusammenfassung in Abschnitt 4.5.1 und das Fazit in Abschnitt 7.1.

3.3.1 Definition der Messunsicherheit

Die Messunsicherheit drückt die Genauigkeit oder Qualität einer Messung aus und wird nach GUM (1993) international wie folgt definiert:

- **uncertainty (of measurement)**
  parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand

Die deutsche Definition nach DIN 1319-1 (1995), DIN 1319-3 (1996) und DIN 1319-4 (1999) lautet:

- **Messunsicherheit**
  Kennwert, der aus Messungen gewonnen wird und zusammen mit dem Messergebnis zur Kennzeichnung eines Wertebereichs für den wahren Wert der Messgröße dient

Die deutsche Definition ist zwar etwas anders formuliert als die englische, soll aber dasselbe bedeuten. Dies folgt aus einer Anmerkung in DIN 1319-3 (1996) und DIN 1319-4 (1999):

- Die Messunsicherheit ist ein Maß für die Genauigkeit der Messung und kennzeichnet die Streuung oder den Bereich derjenigen Werte, die der Messgröße vernünftigerweise als Schätzwerte für den wahren Wert z u g e w i e s e n werden können. Sie kann auch als ein Maß für die Unkenntnis der Messgröße aufgefasst werden.

Die Betonung liegt hier auf dem Wort „zugewiesen“ („attributed“) um auszudrücken, dass der für die Messung Verantwortliche dies zu tun hat. Denn ein vernünftiger Schätzwert folgt nicht automatisch aus der Messung. Außerdem besagt die Formulierung „associated with“, dass die Messunsicherheit dem Messergebnis b e i g e o r d n e t ist. Sie ist n i c h t die „Messunsicherheit des Messergebnisses“, sondern die „Messunsicherheit (der Messgröße) z u m Messergebnis“. Die Messunsicherheit g e h ö r t zwar zum Messergebnis, sie bezieht sich aber auf die Messgröße (measurand), nicht auf das Messergebnis. Die Messgröße ist es, die unvollständig bekannt, also unsicher ist. Das Messergebnis dagegen liegt vor, ist also sicher bekannt. Die Messunsicherheit wird auch kurz *Unsicherheit* genannt [Uns].

3.3.2 Quantifizierung der Messunsicherheit, Standardunsicherheit

Beim physikalischen Messen besteht die Aufgabe, den gerade tatsächlich vorliegenden *„wahren“ Wert* einer physikalische Größe, einer *Messgröße*, z.B. der Geschwindigkeit eines gerade vorbeifahrenden Fahrzeugs, zu messen, was aber im Allgemeinen nur

unsicher möglich ist. Auch hier kann zunächst nur aus der Information, die durch Messung gewonnen wird oder anderweitig gegeben ist, eine Wahrscheinlichkeitsaussage im Bayes'schen Sinne über die möglichen Werte der Messgröße X, d.h. eine Wahrscheinlichkeitsverteilung, aufgestellt werden. (Genauer bezeichnet X eine der Messgröße zugeordnete Zufallsvariable, den Schätzer.) Der Erwartungswert $\mathrm{E}\,X$ dieser Verteilung wird nun als bester *Schätzwert* der Messgröße definiert und *Messergebnis* genannt. Er wird deshalb als der beste aller möglichen Schätzwerte x angesehen, weil für $x = \mathrm{E}\,X$ die nichtzentrale Varianz $\mathrm{E}\,(X - x)^2$ als Maß für die Streuung der möglichen Werte von X um einen Schätzwert x am kleinsten ist. Dann ist $\mathrm{E}\,(X - x)^2 = \mathrm{Var}\,(X)$. Die Wurzel aus dieser Varianz ist die Standardabweichung $\sqrt{\mathrm{Var}\,(X)}$, ein in der Statistik gebräuchliches, anschaulicheres Maß für die Streuung der möglichen Werte von X um den Erwartungswert $x = \mathrm{E}\,X$, das Messergebnis.

Diese Standardabweichung der Verteilung wird nun verwendet, um die Messunsicherheit zu quantifizieren. Sie dient auch dazu, den in der deutschen Definition genannten Wertebereich oder die in der englischen Definition erwähnte Streuung (dispersion) der vernünftigen Schätzwerte (values) der Messgröße nummerisch festzulegen [Uns]. Die Standardabweichung als quantitatives Maß für die Messunsicherheit wird *Standardmessunsicherheit* genannt, kürzer auch *Standardunsicherheit* (*standard uncertainty*) oder einfach *Unsicherheit*, wenn klar ist, dass die Standardmessunsicherheit gemeint ist. Die Standardmessunsicherheit einer Messgröße X zum Messergebnis x wird mit $u(x)$ bezeichnet. Es wird also $u(x) = \sqrt{\mathrm{Var}\,(X)}$ gesetzt. Dann sind $x - u(x)$ und $x + u(x)$ die Grenzen des Bereichs derjenigen Werte, die als vernünftige Schätzwerte der Messgröße neben dem besten Schätzwert x infrage kommen. Dieser Bereich wird im Folgenden *Unsicherheitsbereich* genannt. Er darf nicht mit einem Vertrauensintervall der Statistik verwechselt werden (Abschnitt 3.4.4). Der Quotient $u(x)/x$ wird *relative Standardunsicherheit* genannt.

Es seien nun Merkmale als Messgrößen interpretiert. Dann können entsprechend auf einfache Weise nach Gleichung (6) der Erwartungwert $\mathrm{E}\,Z$ als bester Schätzwert z eines zusammengesetzten Merkmals Z – im Beispiel von Abschnitt 3.2.3 die Gesamtkompetenz eines Teilnehmers – sowie nach Gleichung (7) die Standardabweichung $\sqrt{\mathrm{Var}\,(Z)}$ von Z als die zu diesem besten Schätzwert z gehörende Standardunsicherheit $u(z)$ gebildet werden. z und $u(z)$ gehören als Ergebnis für Z immer zusammen. Es sind dann weiterhin $z - u(z)$ und $z + u(z)$ die Grenzen des Unsicherheitsbereichs des Merkmals Z, also des Bereichs der vernünftigen Schätzwerte für Z. (Der Erwartungswert $\mathrm{E}\,Z$ im Beispiel ist im Allgemeinen keine ganze Zahl, was hier unbeachtet bleiben kann. Siehe Abschnitt 3.4)

Obwohl der Erwartungswert, die Varianz und die Standardabweichung zu einem zusammengesetzten Merkmal auf dessen Wahrscheinlichkeitsverteilung beruhen, können sie berechnet werden, ohne die Verteilung zu kennen (Abschnitt 3.4.4). Deshalb kann auf die Berechnung der Verteilung verzichtet werden, auch weil die Verteilung selbst kaum je benötigt wird und ihre Berechnung in allgemeineren Fällen sehr aufwändig sein kann.

Nachdem zur Quantifizierung der Messunsicherheit die Standardunsicherheit $u(x)$ mit der Standardabweichung identifiziert worden ist, stellt sich nunmehr die Frage, wie die Standardabweichung aus vorliegender Information zu berechnen ist. Diese Frage wird in Abschnitt 3.4.1 untersucht.

3.3.3 Kritischer Vergleich zweier Merkmalsergebnisse

Nach der Vorbereitung von Abschnitt 3.3.2 ist es nun möglich, die Ergebnisse z_1 und z_2 zu zwei gleichartigen Merkmalen Z_1 und Z_2 – im Beispiel von Abschnitt 3.2.3 die Gesamtkompetenzen zweier Teilnehmer 1 und 2 – mit Hilfe der Standardunsicherheiten $u(z_1)$ und $u(z_2)$ kritisch miteinander zu vergleichen. Den Messgrößen beim physikalischen Messen entsprechend, unterscheiden die Ergebnisse sich dann signifikant voneinander, wenn

$$|z_1 - z_2| > k\sqrt{u^2(z_1) + u^2(z_2)} \tag{8}$$

wobei k ein zu vereinbarender Faktor zwischen 1 und 3 ist (Weise und Wöger, 1994, 1999; siehe auch Abschnitt 3.4.4). Oft wird $k = 2$ gewählt (siehe dazu weiter unten). Auch die Wahl $k = \sqrt{2} = 1{,}414$ kann zweckmäßig sein, weil immer $u(z_1) + u(z_2) \leq \sqrt{2} \cdot \sqrt{u^2(z_1) + u^2(z_2)}$. Wenn dann die einfachere Bedingung

$$|z_1 - z_2| > u(z_1) + u(z_2) \tag{9}$$

n i c h t erfüllt ist, ist auch die Bedingung nach Gleichung (8) mit $k = \sqrt{2}$ nicht erfüllt. Die Ergebnisse z_1 und z_2 der Merkmale Z_1 und Z_2 unterscheiden sich also in diesem Fall dann nicht signifikant voneinander, wenn ihre Differenz dem Betrag nach nicht größer als die Summe der Standardunsicherheiten der Merkmale ist. Die Bedingung nach Gleichung (9) lässt sich auch anders interpretieren: Ist sie erfüllt, so überlappen sich die Unsicherheitsbereiche mit den Grenzen $z_1 - u(z_1)$ und $z_1 + u(z_1)$ bzw. $z_2 - u(z_2)$ und $z_2 + u(z_2)$ nicht. Das sind die Bereiche der vernünftigen Schätzwerte für Z_1 bzw. Z_2. Es gibt dann also keinen gemeinsamen vernünftigen Schätzwert der beiden Merkmale, was deren Ergebnisse z_1 und z_2 als signifikant unterschiedlich ausweist. Ist die Bedingung nach Gleichung (9) erfüllt, muss dies nicht in allen Fällen auch für

die Bedingung nach Gleichung (8) gelten. Es kann also vorkommem, dass Ergebnisse zwar nach Gleichung (9), nicht aber nach Gleichung (8) als signifikant unterschiedlich ausgewiesen werden. Im Folgenden wird für Signifikanzprüfungen die einfachere und anschaulichere Bedingung nach Gleichung (9) benutzt, wozu die Wahl $k = \sqrt{2}$ gehört.

Die vorstehenden Betrachtungen begründen einen statistischen Test für die Hypothese, dass die Merkmale Z_1 und Z_2 übereinstimmen. Die Hypothese wird verworfen, wenn die Bedingungen nach den Gleichungen (8) oder (9) erfüllt sind. Sie könnte auf diese Weise fälschlich verworfen werden, wenn sie tatsächlich richtig ist. Oft wird nach der Wahrscheinlichkeit für diesen *Fehler erster Art* gefragt. Sie beträgt höchstens etwa 5 % für $k = 2$ und höchstens etwa 16 % für $k = \sqrt{2}$, wenn eine Normalverteilung für $Z_1 - Z_2$ in beiden Statistiken zugrunde gelegt wird. Dies ist eine Näherung. Damit sie zulässig ist, sollten Z_1 und Z_2 aus mehreren (mindestens drei) Merkmalen zusammengesetzt sein. (Siehe auch die Abschnitte 3.4.4 und 4.5.1 zum Thema Vertrauensintervall.)

3.4 Verallgemeinerung zum Assessment Center

3.4.1 Ein einzelnes Merkmal

Zum Zweck einer Personalauswahl werde ein Merkmal X eines Teilnehmers durch m Beobachter in n unterschiedlichen Übungen eines AC bewertet. Dafür wird eine Bewertungsskala benötigt. Das Merkmal wird zunächst aufgefasst als ein kontinuierliches quantitatives Merkmal mit einer zu ermittelnden Ausprägung beim Teilnehmer, die einem Wert t zwischen gegebenen Grenzen a und b auf einer geeignet gewählten kontinuierlichen Skala entspricht. Dem Merkmal wird eine Zufallsvariable (Schätzer) zugeordnet, die der Einfachheit halber ebenfalls mit X bezeichnet wird und die möglichen Werte t zwischen den gegebenen Grenzen a und b besitzt. Das Intervall von a bis b wird in N gleich lange Teilintervalle unterteilt. Diese werden mit $i = 1$ bis N nummeriert. Die Grenzen des Teilintervalls i sind t_{i-1} und t_i ($t_0 = a$; $t_N = b$; $t_i - t_{i-1} = (b - a)/N$). Die Skala sollte so gewählt werden, dass den Teilintervallen ohne Vorliegen irgendeiner Information über den Teilnehmer jeweils die gleiche Wahrscheinlichkeit $1/N$ zugeordnet werden kann. (Eine Verallgemeinerung auf ungleich lange Teilintervalle und unterschiedliche gegebene Wahrscheinlichkeiten ist zwar möglich, für die Praxis aber wenig sinnvoll.) Die Beobachter werden gebeten, ihre jeweilige unsichere Bewertung des Merkmals aufgrund des gezeigten Verhaltens des Teilnehmers in jeder Übung danach durch Angabe der minimalen Nummer j und der maximalen Nummer k der Teilintervalle, auf die sie setzen würden, auszudrücken. Es darf $j = k$ sein. Also ist $j \leq k$. Die Bewertungen der Beobachter liegen auf diese Weise auf einer gestuften Bewertungsskala von 1 bis N mit der Stufenhöhe 1.

Die Aussage A eines Beobachters: „Ich würde nur auf die Teilintervalle j bis k setzen“ führt dann bei Anwendung des Bernoulli'schen Prinzips, gleich möglichen Werten t die gleiche Wahrscheinlichkeit zuzuordnen, zur Wahrscheinlichkeitsdichte

$$p(t|A) = \frac{1}{t_k - t_{j-1}} \qquad (t_{j-1} \le t \le t_k) \tag{10}$$

mit $p(t|A) = 0$ für sonstige Werte t. Hier sind t_{j-1} die untere Grenze des Teilintervalls j und t_k die obere Grenze des Teilintervalls k, die zur Aussage A gehören. Wird beispielsweise eine Skala mit der Stufenhöhe 1 gewählt und entspricht die Nummer i der Mitte des Teilintervalls i, so sind $t_{j-1} = j - 1/2$ und $t_k = k + 1/2$. Gleichung (10) stellt eine Rechteckverteilung für die Zufallsvariable X dar mit der Breite $t_k - t_{j-1}$ und mit

$$\mathsf{E}\, X|A = (t_k + t_{j-1})/2 \qquad \mathsf{E}\, X^2|A = (t_k^2 + t_k t_{j-1} + t_{j-1}^2)/3 \tag{11}$$

Insgesamt gibt es $M = m \cdot n$ Aussagen der m Beobachter in den n Übungen zum Merkmal X, denen jeweils wieder die gleiche Wahrscheinlichkeit $P(A) = 1/M$ zugewiesen wird, eine von ihnen zufällig auszuwählen. Es kann jedoch sein, dass die Übungen bezüglich X unterschiedlich aussagekräftig sind. Man wird ihnen dann dementsprechend unterschiedliche Wahrscheinlichkeiten $P(A)$ (Gewichte) zuordnen. Der Entwicklungssatz entsprechend Gleichung (1) ergibt nun als Gesamtwahrscheinlichkeitsdichte für das Merkmal X:

$$p(t) = \sum_{A=1}^{M} p(t|A) P(A) \tag{12}$$

mit

$$x = \mathsf{E}\, X = \sum_{A=1}^{M} (\mathsf{E}\, X|A) P(A) \qquad \mathsf{E}\, X^2 = \sum_{A=1}^{M} (\mathsf{E}\, X^2|A) P(A) \tag{13}$$

$$u^2(x) = \mathsf{Var}\,(X) = \mathsf{E}\, X^2 - (\mathsf{E}\, X)^2 = \mathsf{E}\, X^2 - x^2 \tag{14}$$

Wieder ist der Erwartungswert x der beste Schätzwert für die Ausprägung des Merkmals X und $u(x) = \sqrt{\mathsf{Var}\,(X)}$ ist die zum besten Schätzwert x gehörende Standardunsicherheit des Merkmals X.

Hier sind die Beobachter angehalten, ihrer Unsicherheit in der Einschätzung eines Merkmals durch Angabe einer minimalen und einer maximalen Bewertung, die beide gleich sein dürfen, auszudrücken. Dies scheint im Hinblick auf die Praxis des AC die

einfachste und gangbarste Möglichkeit der Informationsgewinnung für die Berechnung der Standardunsicherheit zu sein. Nach Abschnitt 3.2.2 entfällt auf diese Weise bei mehreren Merkmalen zu demselben Teilnehmer auch die Korrelation. Es gibt natürlich noch andere Möglichkeiten der Bewertung und auch der Skalierung für die Bewertung, die jedoch weniger praktikabel erscheinen (Abschnitt 4.3.3).

Betont wird, dass sich X nicht unbedingt wie hier auf das Merkmal eines Teilnehmers zu beziehen braucht. X kann auch Schätzer zu einem beliebigen Sachverhalt sein.

3.4.2 Vergleich mit der konventionellen Statistik

Der Unterschied zur konventionellen Statistik lässt sich hier verdeutlichen: Es sei angenommen, dass sich alle Beobachter ihrer Bewertungen sehr sicher sind und die Aussagen deshalb alle in der Form „Ich würde nur auf Teilintervall k setzen“ vorliegen. Weiterhin sei die Anzahl N der Teilintervalle so groß, dass der Unterschied zwischen t_k und t_{k-1} vernachlässigt werden kann (hier ist $j = k$). Dann gelten nach Gleichung (11) $\mathrm{E}\,X|A = t_k$ und $\mathrm{E}\,X^2|A = t_k^2$ sowie mit $P(A) = 1/M$ nach den Gleichungen (13) und (14)

$$x = \mathrm{E}\,X = \frac{1}{M}\sum_{A=1}^{M} t_k = \bar{t} \qquad \mathrm{E}\,X^2 = \frac{1}{M}\sum_{A=1}^{M} t_k^2 = \overline{t^2} \tag{15}$$

$$u^2(x) = \mathrm{Var}\,(X) = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \overline{t^2} - \bar{t}^2 = \frac{1}{M}\sum_{A=1}^{M}(t_k - \bar{t})^2 \tag{16}$$

Bei den Summen ist zu beachten, dass zu jedem A ein eigenes t_k gehört.

In der konvertionellen Statistik ist der Mittelwert $\bar{t}$ der vorgelegten Bewertungen ein erwartungstreuer Schätzwert für den Erwartungswert der Häufigkeitsverteilung dieser Bewertungen, und die rechte Seite von Gleichung (16), die empirische Varianz der Bewertungen, ist ein asymptotisch erwartungstreuer Schätzwert für die Varianz der Häufigkeitsverteilung. Ein kleiner Unterschied ist nur, dass in der konventionellen Statistik im Nenner von Gleichung (16) meist $M - 1$ statt M steht (Abschnitt 3.5.1), damit der Schätzwert erwartungstreu ist, nicht nur asymptotisch erwartungstreu. Das spielt aber bei großer Anzahl M keine Rolle. Abgesehen von der unterschiedlichen Interpretation, stimmen also im Fall sehr genauer Aussagen vieler Beobachter die Ergebnisse der Bayes'schen und der konventionellen Statistik asymptotisch, d.h. im Grenzfall unbeschränkt verfügbaren Datenmaterials und bei analogem Vorgehen, nummerisch überein. Den Begriff Erwartungstreue gibt es nur in der konventionellen

Statistik. Er bedeutet, dass der Erwartungswert eines Schätzers für einen gesuchten Parameter einer Häufigkeitsverteilung mit diesem Parameter übereinstimmt.

3.4.3 Zusammensetzung mehrerer Merkmale

Es werden nun alle L Merkmale X_l $(l = 1, \ldots, L)$ eines Teilnehmers betrachtet, die im AC zu bewerten sind. Für alle Merkmale soll dieselbe Bewertungsskala wie in Abschnitt 3.4.1 beschrieben verwendet werden. Das aus den Merkmalen X_l zusammengesetzte Gesamtmerkmal Z, z.B. die Gesamtkompetenz des Teilnehmers, die für eine Personalauswahl benötigt wird, wird zunächst so angesetzt:

$$Z = \sum_{l=1}^{L} G_l \, X_l \tag{17}$$

Die $G_l > 0$ sind Gewichte zu den Merkmalen X_l, die vor der Durchführung des AC festzulegen sind (Abschnitt 4.3.1). (Im Beispiel $Z = X + Y$ von Abschnitt 3.2.2 sind mit $X = X_1$ und $Y = X_2$ hier $L = 2$ und $G_1 = G_2 = 1$.) Die Gewichte bestimmen, wie wichtig die zugehörigen Merkmale z.B. für die Personalauswahl sind. Durch nachträgliche Änderung der Gewichte lässt sich auch untersuchen, inwieweit einzelne Merkmale die Personalauswahl beeinflussen. Für Z gilt dieselbe Bewertungsskala wie für die X_l, wenn die Gewichte normiert sind, d.h. ihre Summe gleich 1 ist. Um das zu erreichen, wird die rechte Seite von Gleichung (17) durch die Summe der Gewichte G_l geteilt. Das ergibt

$$Z = \frac{\sum_{l=1}^{L} G_l \, X_l}{\sum_{j=1}^{L} G_j} = \sum_{l=1}^{L} H_l \, X_l \qquad \left(H_l = \frac{G_l}{\sum_{j=1}^{L} G_j} \right) \tag{18}$$

Die normierten Gewichte H_l werden demnach aus den festgelegten Gewichten G_l so berechnet, dass ihre Summe gleich 1 ist.

Sind nun Schätzwerte $x_l = \mathrm{E}\, X_l$ der Merkmale X_l gegeben und auch die zu den x_l gehörenden Standardunsicherheiten $u(x_l)$ der Merkmale mit $u^2(x_l) = \mathrm{Var}\,(X_l)$, so gelten in Verallgemeinerung der Gleichungen (6) und (7) bei genau – d.h. ohne Unsicherheit – festgelegten Gewichten H_l für den besten Schätzwert $z = \mathrm{E}\, Z$ und die zu z gehörende Standardunsicherheit $u(z)$ des Merkmals Z nach [Uns]:

$$z = \sum_{l=1}^{L} H_l \, x_l \tag{19}$$

$$u^2(z) = \sum_{l=1}^{L} H_l^2 \, u^2(x_l) \tag{20}$$

Für die Gültigkeit von Gleichung (20) muss vorausgesetzt werden, dass alle zur Gleichung beitragenden Aussagen unabhängig voneinander entstanden sind, damit die Merkmale X_l zu demselben Teilnehmer nicht miteinander korrelieren. Diese wichtige Voraussetzung wird in Abschnitt 3.5.2 untersucht.

3.4.4 Fortpflanzung von Unsicherheiten

Komplizierter wird es, wenn auch gegebene Unsicherheiten der Gewichte G_l berücksichtigt werden sollen. Denn auch die Gewichte G_l sind unsicher, lassen sich in der Praxis selten genau festlegen. Nach Gleichung (18) ist

$$Z = F(G_1, \ldots, G_L, X_1, \ldots, X_L) \tag{21}$$

eine Funktion der Gewichte G_l und Merkmale X_l. Dann gelten nach [Uns] in linearer Näherung mit gegebenen Werten $g_l = \mathsf{E}\, G_l$ und $u^2(g_l) = \mathsf{Var}\,(G_l)$ für G_l sowie $x_l = \mathsf{E}\, X_l$ und $u^2(x_l) = \mathsf{Var}\,(X_l)$ für X_l

$$z = F(g_1, \ldots, g_L, x_1, \ldots, x_L) \tag{22}$$

$$u^2(z) = \sum_{l=1}^{L} \Big(\frac{\partial F}{\partial G_l}\Big)^2 u^2(g_l) + \sum_{l=1}^{L} \Big(\frac{\partial F}{\partial X_l}\Big)^2 u^2(x_l) \tag{23}$$

Gleichung (23) ist wie Gleichung (20) nur gültig, wenn alle Kovarianzen gleich 0 gesetzt werden können, d.h. Korrelationen generell vernachlässigbar sind. Für die mit den G_l verbundenen Kovarianzen bildet diese wichtige Voraussetzung kein Problem, weil die G_l unabhängig voneinander und unabhängig von den X_l wählbar sind. Die Korrelation der X_l untereinander bedarf jedoch hinsichtlich der Voraussetzung einer genaueren Betrachtung (Abschnitt 3.5.2).

In die partiellen Ableitungen in Gleichung (23) sind die Erwartungswerte g_l und x_l einzusetzen. Da es in der Praxis schwierig ist, die Unsicherheiten der Gewichte überhaupt einzeln festzulegen, wird nur der einfachste Fall betrachtet, dass für a l l e Gewichte lediglich global eine gleiche relative Unsicherheit $u_\mathsf{r} = u(g_l)/g_l$ z.B. in Prozent angegeben werden kann. Nach Bildung der Ableitungen

$$\frac{\partial F}{\partial X_l} = H_l \qquad \frac{\partial F}{\partial G_l} = (X_l - Z)H_l/G_l \tag{24}$$

mit F nach den Gleichungen (18) und (21) und Einsetzen der Werte g_l und x_l folgt schließlich mit den Gleichungen (19), (22) und (23)

$$z = \sum_{l=1}^{L} h_l\, x_l \qquad \Big(h_l = \frac{g_l}{\sum_{j=1}^{L} g_j}\Big) \tag{25}$$

$$u^2(z) = \sum_{l=1}^{L} h_l^2 \Big(u^2(x_l) + (x_l - z)^2 u_\mathrm{r}^2 \Big) \tag{26}$$

Auf die Berechnung der Wahrscheinlichkeitsverteilung des Gesamtmerkmals Z wird verzichtet, nicht nur, weil sie wegen der Abhängigkeit von den anderen Zufallvariablen nach Gleichung (18) sehr aufwändig sein kann, sondern auch, weil sich gerade wegen dieser Abhängigkeit bei nicht zu wenigen Zufallsvariablen in fast allen Fällen näherungsweise eine Normalverteilung mit $\mathrm{E}\, Z = z$ und $\mathrm{Var}\,(Z) = u^2(z)$ ergibt. Außerdem tendiert $u(z)$ mit steigender Anzahl L der zu Z beitragenden Merkmale X_l zu immer kleineren Werten. Das Gesamtmerkmal Z wird dadurch also weniger unsicher, der Aufwand aber steigt.

Aus der Normalverteilung als Näherung für die Verteilung von Z kann auch ein Vertrauensintervall konstruiert werden, das näherungsweise mit einer vorzugebenden Wahrscheinlichkeit α, dem *Vertrauensniveau*, den gesuchten (wahren) Wert t für die Ausprägung des Gesamtmerkmals Z enthält. Hieraus folgt auch Gleichung (8). (Man beachte, dass das Vertrauensniveau in der konventionellen Statistik k e i n e Wahrscheinlichkeit ist.) Das Vertrauensintervall besitzt die Grenzen $z - k \cdot u(z)$ und $z + k \cdot u(z)$ mit dem Quantil k der standardisierten Normalverteilung zur Wahrscheinlichkeit $1 - \alpha/2$. Meist wird $k = 2$ für das Vertrauensniveau $\alpha \approx 95$ % verwendet. Zu betonen ist, dass das Vertrauensintervall keinerlei neue Information zu der Angabe von z und $u(z)$ beiträgt, denn seine Grenzen werden erst aus dieser Angabe berechnet. Das Vertrauensintervall darf nicht mit dem Unsicherheitsbereich, dem Bereich der vernünftigen Schätzwerte von Z, verwechselt werden, der die Grenzen $z - u(z)$ und $z + u(z)$ besitzt (Abschnitte 3.3.1 und 3.3.2).

Die Gleichungen (18) bis (26) gelten nicht nur für das Gesamtmerkmal eines Teilnehmers, sondern entsprechend auch allgemein für jede gewichtete Zusammenfassung von Merkmalen zu einem anderen interessierenden Merkmal. Wenn in den Gleichungen (12) und (13) die Gewichte $P(A)$ der Übungen bezüglich eines Merkmals X vor Durchführung des AC unterschiedlich festgelegt werden und dies nur unsicher möglich ist, müssen auch die damit verbundenen Unsicherheiten berücksichtigt werden. Eine entsprechende Gleichung, die wie Gleichung (26) abgeleitet werden kann, lautet

$$u^2(x) = \mathrm{E}\, X^2 - x^2 + \sum_{A=1}^{M} P^2(A)\, (\mathrm{E}\, X|A - x)^2 {u'_\mathrm{r}}^2 \tag{27}$$

Für die Merkmale X_l sind dann $u^2(x_l)$, entsprechend berechnet nach dieser Gleichung (27), in Gleichung (26) einzusetzen. u'_r ist die globale relative Standardunsicherheit

der Gewichte $P(A)$ und entspricht der globalen relativen Standardunsicherheit u_r der Gewichte G_l. Es kann u'_r gleich u_r gesetzt oder davon unabhängig festgelegt werden. Die Formel zur Unsicherheit von X nach Gleichung (27) enthält vier Beiträge:

1) die Unsicherheiten der Beobachter bezüglich ihrer eigenen Bewertungen durch die Angabe jeweils einer minimalen und maximalen Bewertung in einer Aussage,
2) die Streuung der Bewertungen, wenn mehrere Aussagen vorliegen,
3) den Beitrag, der durch die Stufung der Bewertungsskala entsteht, weil jeder Bewertung ein ganzes Teilintervall von möglichen Werten des Merkmals X entspricht, auch dann, wenn minimale und maximale Bewertung übereinstimmen, und
4) global die Unsicherheit der Gewichte durch u'_r.

Nur der Beitrag 2 basiert auf statistischer Information. Lediglich diese kann in der konventionellen Statistik berücksichtigt werden. Alle genannten Beiträge wirken sich schließlich über Gleichung (26) auf $u(z)$ aus, wobei außerdem u_r Berücksichtigung findet.

3.5 Korrelation

Die Evaluierung in Abschnitt 6.3 wird anhand von Korrelationskoeffizienten durchgeführt. Die Grundlagen dazu werden in Abschnitt 3.5.1 kurz zusammengestellt. Außerdem wird in Abschnitt 3.5.2 der Begriff der teilnehmerbezogenen Korrelation eingeführt. Dieser wird in Abschnitt 6.3.2 für die Untersuchung benötigt, ob die in den Abschnitten 3.4.3 und 3.4.4 erwähnte Voraussetzung als erfüllt angesehen werden kann.

3.5.1 Grundlagen zu Korrelationskoeffizienten

Wenn die gemeinsame Wahrscheinlichkeitsverteilung zweier Zufallsvariablen X und Y vorliegt, ergibt sich der Korrelationskoeffizient von X und Y wie im Folgenden beschrieben. Zuerst werden von X und Y die *Erwartungswerte* $\mathsf{E}\,X$ bzw. $\mathsf{E}\,Y$ sowie die *Varianzen*

$$\mathsf{Var}\,(X) = \mathsf{E}\,(X - \mathsf{E}\,X)^2 \quad \text{bzw.} \quad \mathsf{Var}\,(Y) = \mathsf{E}\,(Y - \mathsf{E}\,Y)^2 \tag{28}$$

und die *Kovarianz*

$$\mathsf{Cov}\,(X, Y) = \mathsf{E}\,(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y) \tag{29}$$

berechnet. Daraus ergibt sich dann der *Korrelationskoeffizient*

$$\varrho(X, Y) = \frac{\mathsf{Cov}\,(X, Y)}{\sqrt{\mathsf{Var}\,(X) \cdot \mathsf{Var}\,(Y)}} \tag{30}$$

Es gelten die Beziehungen $-1 \leq \varrho(X,Y) \leq +1$ und $\varrho(Y,X) = \varrho(X,Y)$. Die Wurzeln $\sigma(X) = \sqrt{\mathrm{Var}\,(X)}$ und $\sigma(Y) = \sqrt{\mathrm{Var}\,(Y)}$ der Varianzen sind die *Standardabweichungen* von X bzw. Y.

In der konventionellen Statistik ist die gemeinsame Wahrscheinlichkeitsverteilung von X und Y nicht bekannt. Es können aber Schätzwerte der Erwartungswerte, Varianzen und Kovarianz gebildet werden, wenn Paare (x_j, y_j) zusammengehörender Daten x_j und y_j zu X bzw. Y vorliegen. m sei die Anzahl der Datenpaare. Schätzwerte der Erwartungswerte $\mathrm{E}\,X$ und $\mathrm{E}\,Y$ sind die *Mittelwerte*

$$\overline{x} = \frac{1}{m}\sum_{j=1}^{m} x_j \quad \text{bzw.} \quad \overline{y} = \frac{1}{m}\sum_{j=1}^{m} y_j \tag{31}$$

Schätzwerte der Varianzen $\mathrm{Var}\,(X)$ und $\mathrm{Var}\,(Y)$ und der Kovarianz $\mathrm{Cov}\,(X,Y)$ sind die *empirischen Varianzen*

$$s_x^2 = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \overline{x})^2 \quad \text{und} \quad s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_j - \overline{y})^2 \tag{32}$$

bzw. die *empirische Kovarianz*

$$s_{xy} = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \overline{x})(y_j - \overline{y}) \tag{33}$$

Aus diesen Schätzwerten wird dann als Schätzwert des Korrelationskoeffizienten $\varrho(X,Y)$ der *empirische Korrelationskoeffizient* wie folgt gebildet:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{j=1}^{m}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{m}(x_j - \overline{x})^2 \cdot \sum_{j=1}^{m}(y_j - \overline{y})^2}} \tag{34}$$

wobei sich der Faktor $1/(m-1)$ herauskürzt. Es gelten die Beziehungen $-1 \leq r_{xy} \leq +1$ und $r_{yx} = r_{xy}$. Die *empirischen Standardabweichungen* s_x und s_y sind Schätzwerte der Standardabweichungen $\sigma(X)$ bzw. $\sigma(Y)$. Sie sind ein Maß für die Streuung der Daten x_j bzw. y_j. Häufig wird auch die *empirische Standardabweichung* $s_{\overline{x}} = s_x/\sqrt{m}$ *des Mittelwertes* $\overline{x}$ gebildet. Sie ist Schätzwert der Standardabweichung einer Zufallsvariablen Z in Form des Mittelwertes $Z = \overline{X} = (1/m)\sum_{j=1}^{m} X_j$ mehrerer anderer Zufallsvariablen X_j, für die je ein Wert x_j vorliegt. Sie ist somit ein Maß für die Streuung der Mittelwerte im Fall, dass der Vorgang zur Gewinnung aller Daten x_j oftmals wiederholt wird. Soweit zur konventionellen Statistik.

In der Bayes'schen Statistik ergibt sich die gemeinsame Wahrscheinlichkeitsverteilung von X und Y aus der vorliegenden Information, zu der auch die Datenpaare (x_j, y_j) gehören können, nach dem in Abschnitt 3.1 beschriebenen Bernoulli'schen Prinzip (oder PME). Sie ist also bekannt und ihre Parameter $\mathrm{E}\,X$, $\mathrm{E}\,Y$, $\mathrm{Var}\,(X)$, $\mathrm{Var}\,(Y)$, $\mathrm{Cov}\,(X,Y)$ und $\varrho(X,Y)$ lassen sich daraus wie oben berechnen. Wenn die Zufallsvariablen X und Y Schätzer für Merkmale sind, sind die Erwartungswerte Schätzwerte der wahren Ausprägungen der Merkmale und die Wurzeln aus den Varianzen liefern die Standardunsicherheiten zu diesen Schätzwerten, außerdem ist der Korrelationskoeffizient ein Schätzwert für die Korrelation der Merkmale.

In beiden Statistiken sind die Zufallsvariablen X und Y und damit die Parameter und in der konventionellen Statistik deren Schätzwerte natürlich von der aktuellen Fragestellung abhängig. Daher bedürfen sie und auch die Datenpaare dahingehend in jedem einzelnen Fall einer genauen Erläuterung. So liegt hinsichtlich der Korrelationskoeffizienten im Fall, dass sie als Validitätsmaß für Merkmale in Übungen benutzt werden sollen, eine ganz andere Fragestellung vor als im Fall der Untersuchung der Korrelation im Rahmen der Berechnung der Unsicherheit von Merkmalen. Dementsprechend handelt es sich in diesen Fällen, auch wenn es sich in beiden um Merkmale zu drehen scheint, um unterschiedliche Korrelationskoeffizienten.

3.5.2 Teilnehmerbezogene Korrelation

Für die Gültigkeit der Gleichungen (20), (23) und (26) muss vorausgesetzt werden, dass alle zu den Merkmalen X_l vorliegenden Aussagen unabhängig voneinander entstanden sind, damit die Merkmale X_l zu demselben Teilnehmer über alle Übungen des AC hinweg nicht miteinander korrelieren. Bei den Aussagen unterschiedlicher Beobachter ist das nicht problematisch, wenn man unterstellt, dass die Beobachter sich nicht gegenseitig informieren. Bei den Aussagen eines Beobachters zu unterschiedlichen Übungen oder Merkmalen lässt sich aber eine Abhängigkeit kaum ganz vermeiden, selbst wenn sich der Beobachter bemüht, seine Aussagen zu vorangegangenen Übungen oder anderen Merkmalen zu vergessen. Allerdings liegt zu einer eventuellen Abhängigkeit von Aussagen im einzelnen AC keinerlei Information vor, lässt sich auch nicht gewinnen und deshalb nicht berücksichtigen (Abschnitt 3.2.2). Es kann auch sein, dass zwei Merkmale voneinander abhängen, weil sie nicht genügend unabhängig definiert sind, d.h. zwar unterschiedlich benannt sind, aber dennoch eine ähnliche oder entgegengesetzte Bedeutung haben. Durch eine geeignete Wahl z.B. kleinerer Gewichte lässt sich erreichen, dass die Merkmale in einem solchen Fall nicht zusammenwirkend dominieren.

Betont wird, dass die Gleichungen (20), (23) und (26) für einen einzelnen Teilnehmer gelten. Das bedeutet, dass nur die Merkmale (genauer die Schätzer) X_l zu

d e m s e l b e n Teilnehmer über alle Übungen hinweg als nicht miteinander korreliert vorausgesetzt werden müssen. Diese Voraussetzung, dass, anders ausgedrückt, die *teilnehmerbezogene Korrelation* vernachlässigt werden darf, ist wesentlich für die Berechnung der Unsicherheit eines zusammengesetzten Merkmals nach Gleichung (20) oder (26). Sie ist nicht erforderlich für Merkmale zu unterschiedlichen Teilnehmern und Übungen und auch nicht für Zufallsvariablen zur Untersuchung der Konstruktvalidität des AC, wobei Information von allen Teilnehmern und nur jeweils aus einer Übung oder einem Übungspaar herangezogen wird (siehe unten und Abschnitte 6.2 und 6.3).

Die Gleichungen (20), (23) und (26) für die Berechnung von Standardunsicherheiten fußen letztlich auf der einfachsten Gleichung (7) für ein aus zwei Merkmalen X und Y zusammengesetztes Merkmal Z eines Teilnehmers unter der Voraussetzung, dass sich die hier vorkommende teilnehmerbezogene Kovarianz vernachlässigen lässt, also $\mathrm{Cov}\,(X, Y) = 0$ gesetzt werden kann. Im Allgemeinen ist das bei dem einzelnen Teilnehmer sicher nicht ohne Weiteres möglich. Unter der Vermutung jedoch, dass bei diesem Teilnehmer die Kovarianz nur zufallsbedingt von null verschieden ist, sollte diese wenigstens dann näherungsweise gleich null sein, wenn gemeinsame Aussagen zu b e i d e n Merkmalen X und Y des Teilnehmers von sehr v i e l e n Beobachtern vorliegen. Aber das lässt sich schwer prüfen, weil die beiden Merkmale des Teilnehmers im AC meist nur von sehr wenigen Beobachtern bewertet werden, oft sogar nur von einem einzigen Beobachter oder aber von unterschiedlichen Beobachtern. Wenn das AC allerdings auf genügend viele Teilnehmer angewendet wird, sollte, wenn die obige Vermutung zutrifft, die Kovarianz wenigstens im Mittel über alle Teilnehmer näherungsweise gleich null sein.

Wird dementsprechend Gleichung (7) über alle Teilnehmer gemittelt, folgt

$$\overline{\mathrm{Var}\,(Z)} = \overline{\mathrm{Var}\,(X)} + \overline{\mathrm{Var}\,(Y)} + 2\,\overline{\mathrm{Cov}\,(X, Y)} \tag{35}$$

woraus sich analog zu Gleichung (30) der *teilnehmerbezogene Korrelationskoeffizient*

$$\delta(X, Y) = \frac{\overline{\mathrm{Cov}\,(X, Y)}}{\sqrt{\overline{\mathrm{Var}\,(X)} \cdot \overline{\mathrm{Var}\,(Y)}}} \tag{36}$$

bilden lässt. Wenn dieser bei sehr vielen Teilnehmern nicht signifikant von null abweicht, ist die gemittelte Kovarianz im Zähler des Quotienten von Gleichung (36) gegenüber der Wurzel im Nenner vernachlässigbar. Dann sollte auch bei einem einzelnen Teilnehmer die teilnehmerbezogene Kovarianz als nur zufallsbedingt vernachlässigt werden dürfen. Das wird in Abschnitt 6.3.3 im Rahmen der Evaluierung untersucht.

Hier soll noch an einem Beispiel dargelegt werden, was genau unter teilnehmerbezogener Korrelation und deren Zufallsbedingtheit zu verstehen ist.

In Bild 3.1 sind als Beispiel zu drei Teilnehmern (T) jeweils drei Rechtecke gezeichnet. Eine gestufte Bewertungsskala von 1 bis 10 mit der Stufenhöhe 1 ist zugrunde gelegt. Jedes Rechteck stellt die Aussage ein und d e s s e l b e n Beobachters bezüglich z w e i e r Merkmale X und Y eines Teilnehmers dar in der Form „Ich würde bei X nur auf die Teilintervalle i bis j und bei Y nur auf die Teilintervalle k bis l setzen“. Aufgrund allein dieser Aussage ist dann nach dem Bernoulli'schen Prinzip die gemeinsame Wahrscheinlichkeitsdichte der Merkmale X und Y innerhalb des Rechtecks konstant und außerhalb gleich null. Durch Überlagerung der Dichten der einzelnen Rechtecke aufgrund der Aussagen der Beobachter zu demselben Teilnehmer und aus allen Übungen entsteht die gesamte gemeinsame Wahrscheinlichkeitsdichte der Merkmale X und Y zu diesem Teilnehmer. Das Paar $x = \mathrm{E}\,X$ und $y = \mathrm{E}\,Y$ der Erwartungswerte zum Teilnehmer bildet den in Bild 3.1 mit $\times$ markierten Punkt. Bezüglich dieser teilnehmereigenen Erwartungswerte ist die teilnehmerbezogene Kovarianz zu berechnen. Die Wurzeln aus den entsprechend berechneten Varianzen sind die Standardunsicherheiten $u(x)$ und $u(y)$. Zu beachten ist, dass bei den Erwartungswerten und Varianzen alle Aussagen zu X des Teilnehmers und getrennt davon zu Y heranzuziehen sind, bei der Kovarianz jedoch nur Paare von Aussagen zu X und Y, die von demselben Beobachter stammen. Anschließend sind die Varianzen und die Kovarianz über alle Teilnehmer zu mitteln und der teilnehmerbezogene Korrelationkoeffizient $\delta(X, Y)$ nach Gleichung (36) zu bilden.

Nach Abschnitt 3.2.2 ist die teilnehmerbezogene Kovarianz und damit auch der teilnehmerbezogene Korrelationskoeffizient bei nur einer einzigen vorliegenden Aussage pro Teilnehmer exakt gleich null. Sie sind es auch dann, wenn die Rechtecke zu den Aussagen mehrerer Beobachter dasselbe Zentrum aufweisen. Sie sind ebenso exakt gleich null im Fall, dass gar keine Aussagen eines Beobachters zu beiden Merkmalen vorliegen, d.h. die beiden Merkmale von unterschiedlichen Beobachtern bewertet wurden, sonst nur in Sonderfällen, im Allgemeinen aber nicht. Bei sehr vielen Aussagen zu demselben Teilnehmer kann man jedoch vermuten, dass die Rechtecke zu diesen Aussagen „rein zufällig“ um den Punkt verteilt liegen, der das Paar der Erwartungswerte zu diesem Teilnehmer darstellt. Das aber bedeutet, dass der Betrag des teilnehmerbezogenen Korrelationskoeffizienten sehr klein ist und deshalb die Korrelation vernachlässigt werden kann. Das gilt in gleicher Weise bei jedem einzelnen Teilnehmer bezüglich seiner eigenen Erwartungswerte. Daraus folgt ein Prüfverfahren dafür bei vielen Teilnehmern, aber nur wenigen Aussagen zu jedem von ihnen: Die Rechtecke zu jedem einzelnen Teilnehmer werden, ohne ihre gegenseitige Lage zu verändern oder sie zu verdrehen,

so verschoben, dass schließlich alle Punkte zu den Paaren der Erwartungswerte der Teilnehmer zusammenfallen (Bild 3.1b). Es wird also so getan, als gehörten alle Aussagen zu demselben „mittleren“ Teilnehmer. Damit wird der auf diesen Teilnehmer bezogene Korrelationskoeffizient neu berechnet. Das geschieht durch Gleichung (36). Ist sein Betrag klein genug, kann die wesentliche Voraussetzung als erfüllt angesehen werden.

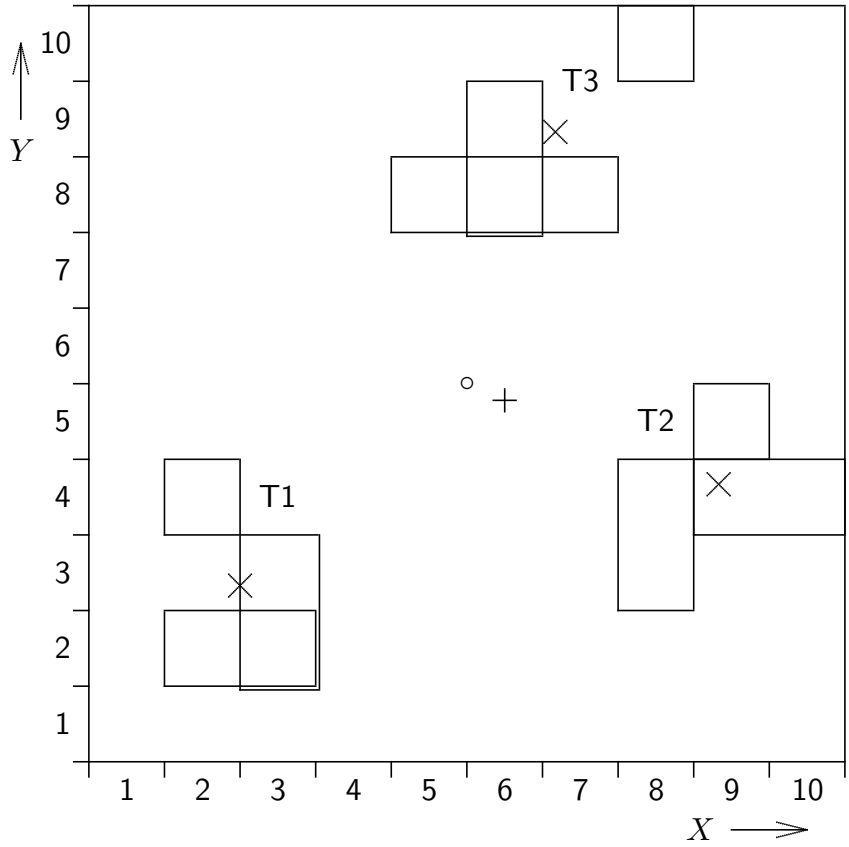
Von der teilnehmerbezogenen Korrelation ist die Korrelation von X und Y zu unterscheiden, die in Abschnitt 6.3 zur Beurteilung der Konstruktvalidität benutzt wird. Bei dieser Korrelation bezieht sich jedes der Merkmale X und Y nur auf eine einzige Übung, und es sind alle diesbezüglichen Aussagen heranzuziehen. Auf den Paaren dieser Aussagen auch unterschiedlicher Beobachter, aber jeweils zu demselben Teilnehmer, beruhen dann die in Bild 3.1a dargestellten Rechtecke. Aus diesen u n v e r s c h o b e n e n Rechtecken zu allen Teilnehmern ist die Korrelation bezüglich der Gesamterwartungswerte zu berechnen. Der dazu gehörende Punkt ist in Bild 3.1a mit $+$ markiert und sollte bei vielen Teilnehmern nahe beim Zentrum $\circ$ des Bewertungsquadrats liegen. Der so berechnete Korrelationskoeffizient kann auch bei sehr vielen Aussagen und Teilnehmern im Allgemeinen stark von null abweichen. Im Gegensatz zum teilnehmerbezogenen Korrelationskoeffizienten hängt er nicht davon ab, ob die Aussagen zu X und Y von denselben oder unterschiedlichen Beobachtern stammen.

Bild 3.1: Beispiel zur Erläuterung der teilnehmerbezogenen Korrelation

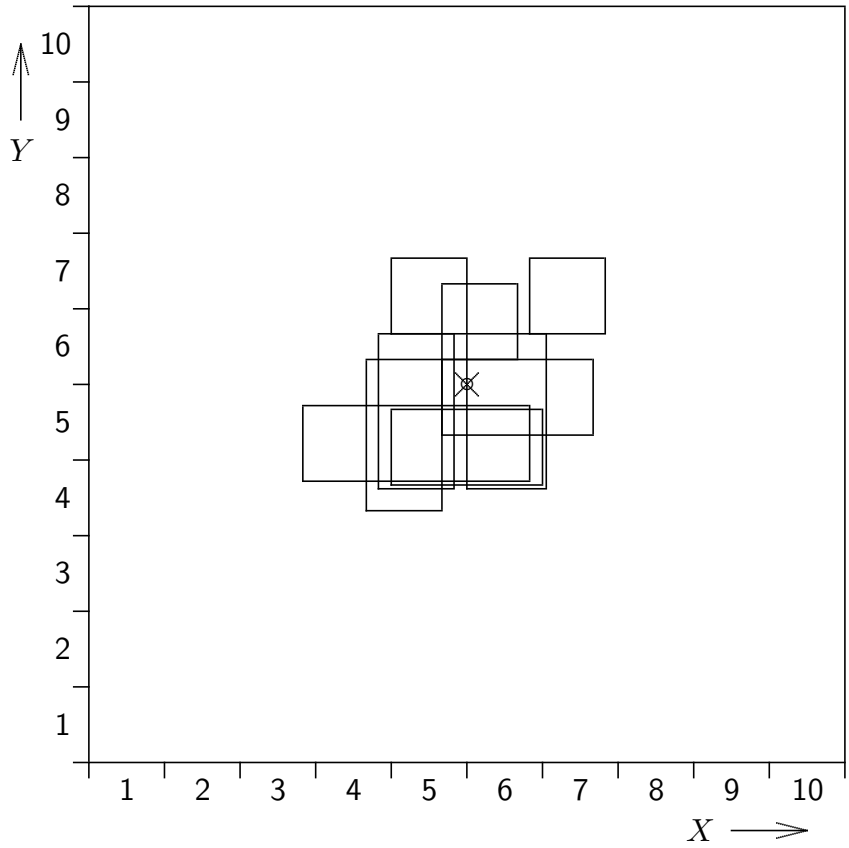
(Bild siehe folgende Seite)

Jedes Rechteck stellt eine Aussage eines Beobachters zu zwei Merkmalen X und Y dar. Zu drei Teilnehmern (T) sind je drei solcher Rechtecke gezeichnet, z.B. bedeutet das linke Rechteck bei T2 die Bewertung 8 für X und 3 bis 4 für Y. Eine gestufte Bewertungsskala von 1 bis 10 mit der Stufenhöhe 1 ist zugrunde gelegt. Es besteht im Allgemeinen Korrelation bezüglich des Punktes $+$ zum Paar der Gesamterwartungswerte der Verteilung aller Rechtecke (Teilbild a). Werden jedoch die Punkte $\times$ zu den Paaren der teilnehmerbezogenen Erwartungswerte mit den jeweils zugehörigen Rechtecken so verschoben, dass sie z.B. im Zentrum $\circ$ des Bewertungsquadrats zusammenfallen, sollte die sich dann ergebende teilnehmerbezogene Korrelation bei vielen Aussagen oder vielen Teilnehmern verschwinden (Teilbild b).

a)



b)



4 Verfahren zur Auswertung eines Assessment Centers

In diesem Kapitel 4 wird das Verfahren zur Auswertung eines AC unter Berücksichtigung der Unsicherheit entwickelt. Das Verfahren besteht aus mehreren Schritten:

1) Konstruktion des AC im Wesentlichen wie üblich,
2) Aufstellung des Modells der Auswertung, das besagt, welche Merkmale wie gewichtet zusammenzusetzen sind,
3) Datenvorbereitung mit Festlegung der Gewichte, Wahl der Bewertungsskala und Erfassung der Bewertungen,
4) Berechnung der Teilnehmer-Ergebnisse mit den zugehörigen Unsicherheiten.

Außerdem wird beschrieben, wie diese Unsicherheiten zu interpretieren sind und wie ein Computer-Programmsystem für die Durchführung des Verfahrens in der Praxis gestaltet werden sollte.

4.1 Konstruktion eines Assessment Centers

In diesem Abschnitt geht es um die Festlegung der Merkmale und Übungen eines AC. Weiterhin werden die Operationalisierung von Merkmalen, z.B. durch Verhaltensanker, sowie die Zusammenfassung von Merkmalen in Merkmalsgruppen besprochen.

Ein AC, das nach dem Verfahren in dieser Arbeit ausgewertet werden soll, ist hinsichtlich der zu bewertenden Merkmale und der hierfür abzuhaltenden Übungen sowie bezüglich seiner praktischen Durchführung im Wesentlichen wie üblich zu konstruieren bzw. zu planen. Anforderungen dahingehend an Verfahren bei berufsbezogenen Eignungsbeurteilungen – wozu auch das AC und seine Übungen gehören – sowie an deren Einsatz sind in DIN 33430 (2002) festgelegt. Als erster Schritt bei der Konstruktion eines AC ist eine gründliche Anforderungsanalyse durchzuführen. Einen Überblick über verschiedene Methoden dazu gibt Krumbach (1999). Grundsätzliche Aspekte zur praktischen Durchführung finden sich u.a. beim Arbeitskreis Assessment-Center (1992), bei Fisseni und Fennekels (1995), Obermann (1992) und Sarges (1996). Aufbauend auf der Anforderungsanalyse sind die Übungen (Jochmann, 1995; Lehment, 1999) und die Testverfahren (Brickenkamp, 1997; Hossiep, Paschen und Mühlhaus, 2000) zu konstruieren oder auszuwählen. Vor Durchführung des AC sollte eine Beobachterschulung stattfinden, bei der die Beobachter mit den Übungen und Merkmalen sowie mit der Beobachtungsmethode vertraut gemacht werden (Drees, 1994; Niermeyer, 1999). Im

Anschluss an das AC sollte mit jedem Teilnehmer ein detailiertes Rückmeldegespräch geführt werden (Schubert, 1999). Eine ausführliche Diskussion zur Konstrukt- und Kriteriumsvalidität in Assessment Centern findet sich bei Kleinmann (1997) und Scholz (1994). Aus mehreren umfangreichen Studien leitet Kleinmann praktische Hinweise zur Konstruktion und Durchführung eines AC ab (z.B. hinsichtlich der Zielsetzung – ob Potenzial- oder Auswahl-AC –, Bekanntgabe der Merkmale, Rotation der Beobachter, Art und Weise der Beobachtung, Methoden der Evaluierung). Die Ratschläge aus der Literatur sollten befolgt werden. Sie werden hier nicht diskutiert, dennoch sind einige Bemerkungen zur Festlegung von Merkmalen und Übungen angebracht.

Nach Abschnitt 2.2.5 wird ein Merkmal als zu bewertende Eigenschaft eines Teilnehmers verstanden. Dementsprechend sollte jedes Merkmal einen klar und genau beschriebenen Sachverhalt charakterisieren, der für das Ziel des AC wichtig und aussagekräftig ist und der sich von den Beobachtern in einer Übung auch gut beobachten lässt. Eine Übung ist deshalb so zu gestalten, dass sie zu jedem in dieser Übung zu bewertenden Merkmal genügend Information erbringt, damit die Beobachter eine vernünftige Bewertung abgeben können. Auch sollte die Übung alle wichtigen Aspekte zu jedem dieser Merkmale abdecken (Kleinmann, 1997). Falls das für ein Merkmal nicht vollständig möglich ist, sollten weitere Übungen zu diesem Merkmal vorgesehen werden, die insgesamt alle Aspekte erhellen.

Das Gleiche gilt für Verhaltensanker, den Komponenten eines operationalisierten Merkmals, denn auch diese sind beobachtbare Sachverhalte. Sollen die Verhaltensanker selbst einzeln bewertet werden, können sie deshalb als eigene Merkmale angesehen werden. Es ist jedoch häufig problematisch, für die Verhaltensanker als Merkmale jeweils eine feste Ausprägung vorauszusetzen, wie es nach Abschnitt 2.2.5 erforderlich ist. Deshalb erscheint es besser, alle Bewertungen der Verhaltensanker gleich als Information über das operationalisierte Merkmal aufzufassen, also als direkte Bewertungen dieses Merkmals durch die Beobachter anzusehen. In den Kapiteln 5 und 6 wird die AC-Serie ST auf diese Weise ausgewertet. Es ist zweckmäßig, in jeder Übung ein spezielles Bewertungsformular (Abschnitt 4.3.4) oder eine Verhaltensscheckliste zu benutzen, worauf alle jeweils zu beobachtenden Verhaltensanker aufgeführt sind (Kleinmann, 1997).

Alle zu bewertenden Merkmale sollten möglichst stark differierende Sachverhalte charakterisieren. Die Merkmale sollten also wenig miteinander zu tun haben, unabhängig sein und daher nicht korrelieren. Diese Forderung ist wichtig, ob sie aber erfüllt ist, lässt sich nicht leicht nachweisen, insbesondere nicht in einem singulären AC. Deshalb

sollten die Merkmale bei der Konstruktion des AC dahingehend sehr kritisch betrachtet werden. Sie sollten aus guten Gründen wenigstens näherungsweise als unabhängig vorausgesetzt werden können. Stark voneinander abhängige Merkmale sollten identifiziert, also zu einem einzigen Merkmal zusammengefasst werden oder in ihrem Gewicht gegenüber anderen Merkmalen reduziert werden (Abschnitt 4.3.1).

Zweckmäßig ist es, auch solche Merkmale und Übungen im AC vorzusehen, die bereits auf viele Teilnehmer angewendet worden sind und deren Konstruktvalidität statistisch gesichert worden ist. Dadurch kann die Validität des AC erhöht werden. Neben der Konstruktion maßgeschneiderter Übungen für die Merkmale zu den Fragestellungen der aktuellen Aufgabe sollten deshalb auch Standardmerkmale und Standardübungen herangezogen werden, soweit sie für Aufgabe relevant sind oder an diese leicht angepasst werden können. Darunter sollten sich auch standardisierte Tests befinden (Hossiep, Paschen und Mühlhaus, 2000; Schmidt und Hunter, 2000).

Die vorzusehenden Merkmale sollten alle für das Ziel des AC wesentlichen Sachverhalte abdecken, also vollständig sein. Das ist wichtig, um Fehleinschätzungen oder Aufwand für nachträglich nötige Bewertungen zu vermeiden (Abschnitt 6.1.2). Um den Aufwand für das AC zu begrenzen und Transparenz für die Beobachter zu erhalten, sollten insgesamt höchstens etwa 15 Merkmale bewertet werden, in jeder Übung höchstens etwa fünf. Es sollten nur solche Teilnehmer zugelassen werden, die alle unabdingbaren, also wirklich wichtigen Anforderungen auch tatsächlich voll erfüllen. Die Durchführung eines Vorauswahlverfahrens – ebenfalls ein (einfaches) AC im Sinne dieser Arbeit – ist deshalb zweckmäßig. Die Merkmale oder auch nur das Gesamtmerkmal dieser Vorauswahl können ebenfalls als Merkmale des AC aufgefasst werden und auf diese Weise angemessene Berücksichtigung finden (Abschnitt 4.2). Bei einer Entscheidungsaufgabe brauchen Merkmale, die von vornherein bei allen Teilnehmern im gleichen Maße ausgeprägt sind, nicht beachtet zu werden, da sie zur Entscheidungsfindung nichts beitragen. Mehrere weniger wichtige Merkmale können zu einem Sammelmerkmal zusammengefasst werden.

Oft ist es zweckmäßig, aus zusammengehörenden Merkmalen Merkmalsgruppen zu bilden. Das erleichtert die Übersicht und die Vergabe von Gewichten (Abschnitt 4.3.1). Jedes Merkmal darf aber nur höchstens einer einzigen Merkmalsgruppe angehören (Abschnitt 4.4.1). Ein Merkmal kann auch allein eine Merkmalsgruppe bilden, z.B. das Vorauswahlmerkmal. Dieses Merkmal ist dann mit der Merkmalsgruppe identisch.

Üblich und nahezu unabdingbar für die Übersicht in einem AC ist es, eine Tabelle anzulegen, worin in der ersten Spalte untereinander alle Merkmalsgruppen mit ihren

Merkmalen und diese mit ihren Verhaltensankern, wenn welche zu bewerten sind, aufgeführt werden. Für jede Übung ist eine weitere Spalte der Tabelle vorzusehen. Auf diese Weise entsteht eine Matrix. Darin ist einzutragen, welches Merkmal und welcher Verhaltensanker in welcher Übung zu bewerten ist oder nicht und mit welchem Gewicht diese Bewertung bei der Auswertung zu berücksichtigen ist. Ein Schema einer solchen Matrix der Merkmale und Übungen ist in Tabelle 4.1 dargestellt, zu Beispielen siehe die Tabellen 4.2, 5.2 und 5.3. (Die Gewichte in den Tabellen 4.1 und 4.2 können zunächst unbeachtet bleiben, sie werden in Abschnitt 4.3 behandelt.) Beim Vergleich der Matrix von Tabelle 4.1 mit den Beispielen in den Tabellen 4.2, 5.2 und 5.3 ist zu beachten, dass es nicht auf die äußere Form der Matrix ankommt, sondern auf den logischen Inhalt.

4.2 Aufstellung des Modells der Auswertung

Das in diesem Kapitel 4 zu entwickelnde Auswerteverfahren eines AC, das zusätzlich zu der üblichen Auswertung durch Berücksichtigung der Unsicherheiten zu den Bewertungen der einzelnen Merkmale der Teilnehmer die in Abschnitt 1.4 genannten Vorteile erbringen soll, basiert wie jede Auswertung von Messungen auf einem *Modell der Auswertung* [Uns].

Das Modell beschreibt, wie die verschiedenen Merkmale X_l $(l = 1, \ldots, L)$ eines Teilnehmers zu einem Gesamtmerkmal Z des Teilnehmers gewichtet zusammenzusetzen sind. Dies geschieht nach Gleichung (17). Das Modell lautet danach

$$Z = \sum_{l=1}^{L} H_l \, X_l \tag{37}$$

Das zusammengesetzte Merkmal Z wird also als gewichtetes Mittel der L Merkmale X_l berechnet. Die Gewichte H_l sind erforderlich, weil z.B. für eine Entscheidung die einzelnen beitragenden Merkmale unterschiedlich wichtig sein können. Zu den Gewichten siehe auch Abschnitt 3.4.3, zur Festlegung ihrer Werte Abschnitt 4.3.1. Die Summe der Gewichte H_l muss gleich 1 sein, damit die für die Merkmale X_l verwendete Bewertungsskala auch für Z gilt. Wenn alle Gewichte H_l gleich sind, ist $H_l = 1/L$ und Z ist der arithmetische Mittelwert der X_l.

Im Modell der Auswertung nach Gleichung (37) ist es gleichgültig, ob Z das Gesamtmerkmal aller Merkmale eines Teilnehmers darstellt oder nur ein aus den Merkmalen einer Merkmalsgruppe des Teilnehmers zusammengesetztes Gruppenmerkmal. Im letzteren Fall kann dieses mit anderen solcher Gruppenmerkmale wieder entsprechend

Gleichung (37) zu einem Gesamtmerkmal zusammengesetzt werden, wobei dann die X_l die Gruppenmerkmale sind. Dieser zweistufige Prozess ist oft zweckmäßig. (Auch mehrstufige Prozesse sind möglich, aber weniger praktikabel und daher nicht empfehlenswert.) Die Merkmalsgruppen dürfen aber nicht überlappen, jedes Merkmal darf höchstens einer Merkmalsgruppe angehören. Anderenfalls ist die für die Berechnung der Unsicherheit wichtige und in Abschnitt 3.5.2 behandelte Voraussetzung nicht mehr erfüllt. Denn dann beeinflusst ein Merkmal mindestens zwei Gruppenmerkmale, die dadurch korreliert sind. Es könnten auch, wie es beim Messen in der Physik häufig erforderlich ist [Uns], wesentlich kompliziertere Modelle verwendet werden, auch solche, bei denen Merkmale stark korreliert sind, jedoch erscheint das für die Praxis des AC derzeit, da es erst darauf ankommt, den Begriff der Unsicherheit einzuführen und Anwender damit vertraut zu machen, verfrüht und unnötig.

4.3 Datenvorbereitung

4.3.1 Festlegung der Gewichte

Siehe hierzu auch die Abschnitte 3.4.3 und 3.4.4.

Die Wahl der Werte für Gewichte ist eine Hauptaufgabe bei der Vorbereitung des AC. Es werden einerseits Werte h_l für die Gewichte H_l in Gleichung (37) benötigt, mit denen Merkmale zu zusammengesetzten Merkmalen beitragen, andererseits auch Werte für die Gewichte $P(A)$ der Aussagen A der Beobachter hinsichtlich ihrer Bewertungen. Diese sind für Gleichung (27) erforderlich. Üblicherweise werden gar keine Gewichte festgelegt, was genauer bedeutet, dass implizit gleiche Gewichte für alle Beiträge angenommen werden und das Gewicht gleich null bei einem Merkmal oder Verhaltensanker, das oder der in einer Übung nicht zu bewerten ist. Das ist aber keineswegs immer sinnvoll. Zur Berücksichtigung entscheidungsanalytischer Verfahren und Festlegung von Gewichten bei der Personalbeurteilung siehe auch Jungermann (1995).

Die Werte der Gewichte sind in die Matrix der Merkmale und Übungen einzutragen. Siehe hierzu die Tabellen 4.1 und 4.2. Für die Wahl der Gewichte kann die Beantwortung der folgenden Fragen hilfreich sein:

1) Gewichte der Merkmalsgruppen zueinander: Wie wichtig ist diese Gruppe im Vergleich zu jeder anderen Gruppe? Dies betrifft in der Matrix in Tabelle 4.1 die mit • bezeichneten Stellen in der Spalte „Merkmalsgruppen“.
2) Gewichte der Merkmale zueinander innerhalb einer Merkmalsgruppe: Wie wichtig ist dieses Merkmal im Vergleich zu jedem anderen in derselben Gruppe? Dies betrifft in der Matrix in Tabelle 4.1 die mit ∘ bezeichneten Stellen in der Spalte

„Merkmale“, jedoch nur innerhalb einer Gruppe, d.h. zwischen den langen horizontalen Linien.

3) Gewichte der Übungen zueinander hinsichtlich eines einzelnen Merkmals: Wie a u s s a g e k r ä f t i g ist diese Übung für das gerade betrachtete Merkmal im Vergleich zu jeder anderen Übung? Dies betrifft in der Matrix in Tabelle 4.1 die mit + bezeichneten Stellen in jeder Merkmals z e i l e.
4) Gewichte der Verhaltensanker eines Merkmals zueinander in einer bestimmten Übung: Wie aussagekräftig ist dieser Verhaltensanker des gerade betrachteten Merkmals im Vergleich zu jedem anderen Verhaltensanker desselben Merkmals in derselben Übung? Dies betrifft in der Matrix in Tabelle 4.1 die mit $\times$ bezeichneten Stellen in einer Übungsspalte zwischen den kurzen horizontalen Linien.
5) Gewicht der Vorauswahl: Wie wichtig ist die Vorauswahl im Vergleich zum aktuellen AC? Für die Vorauswahl sind ein oder mehrere eigene Merkmale vorzusehen, die eine eigene Gruppe bilden können (Beispiel siehe Tabelle 4.2).

Die Gewichte brauchen nur im Verhältnis zueinander für die gerade betrachtete Menge von Merkmalsgruppen, Merkmalen, Übungen oder Verhaltensanker festgelegt zu werden, z.B. 2:5:7 bei drei Merkmalen in Gleichung (37). Die Zahlen 2, 5, 7 entsprechen dabei Werten g_l der Gewichte G_l in den Abschnitten 3.4.3 und 3.4.4. Wenn kein Element aus dieser Menge den anderen bevorzugt werden kann oder soll, sind allen Elementen gleiche Gewichte zuzuweisen. Die nötige Normierung der Gewichte auf die Summe 1 sollte das Auswerteprogramm übernehmen (Abschnitt 4.6.3 und Anhang B). Beispielsweise genügt es, die Gewichte der Merkmale nur innerhalb einer Merkmalsgruppe im Verhältnis zueinander zu setzen. Auch die Gewichte der Übungen hinsichtlich eines Merkmals brauchen nur im Verhältnis zueinander entsprechend ihrer Aussagekraft für dieses Merkmal angegeben zu werden. Ebenso ist es ausreichend, die Gewichte der Merkmalsgruppen im Verhältnis zueinander zu wählen. Entsprechend gilt dies für die Verhaltensanker zu einem Merkmal in einer bestimmten Übung. Die Gewichte der Merkmale oder Übungen können auch gleich null sein. Falls aber die Gewichte aller Merkmale in einer Gruppe oder aller Übungen zu einem Merkmal gleich null gesetzt werden, ist auch das Gewicht der Gruppe bzw. des Merkmals gleich null zu setzen, wenn nicht das Ergebnis zu diesem Merkmal schon vorliegt, z.B. aus einer Vorauswahl (Beispiel siehe Tabelle 4.2). Generell ist das Gewicht einer Übung hinsichtlich eines Merkmals oder Verhaltensankers gleich null zu setzen, wenn letztere in der Übung nicht zu bewerten sind. Wenn gar keine Gewichte gegeben sind, sind alle Gewichte zu Merkmalen oder Verhaltensankern in einer Übung, wenn sie zu bewerten

Tabelle 4.1: Schema einer Matrix der Merkmale und Übungen mit Gewichten

Vergabe der Gewichte spaltenweise zueinander:
•: Merkmalsgruppen
○: Merkmale innerhalb der Merkmalsgruppe
×: Verhaltensanker innerhalb des Merkmals und der Übung:
Vergabe der Gewichte zeilenweise zueinander:
+: Übungen bezüglich eines Merkmals

Merkmalsgruppen Merkmale Verhaltensanker	**Gewichte:**	Merkmals- gruppen	Merkmale	Übungen 1	2	3	...
Gruppe 1		•					
Merkmal 1.1			○	+	+	+	
Merkmal 1.2			○	+	+	+	
Merkmal 1.3			○	+	+	+	
...							
Gruppe 2		•					
Merkmal 2.1			○	+	+	+	
Merkmal 2.2			○	+	+	+	
...							
Gruppe 3		•					
Merkmal 3.1			○	+	+	+	
Verhaltensanker 3.1.1				×	×	×	
Verhaltensanker 3.1.2				×	×	×	
Verhaltensanker 3.1.3				×	×	×	
...							
Merkmal 3.2			○	+	+	+	
Verhaltensanker 3.2.1				×	×	×	
Verhaltensanker 3.2.2				×	×	×	
...							
Merkmal 3.3			○	+	+	+	
...							
Gruppe 4		•					
...							
...							

Tabelle 4.2: Beispiel einer Matrix der Merkmale und Übungen mit Gewichten

Übungen:
1: Mitarbeitergespräch
2: Kundengespräch
3: Konzeptentwicklung, Präsentation
4: Kognitiver Test
fett: Vergabe der Gewichte eines Merkmals in den Übungen zeilenweise zueinander (ansonsten spaltenweise Vergabe)

Merkmalsgruppen **Gewichte:** Merkmale Verhaltensanker	Merkmals-gruppen	Merkmale	Übungen 1	2	3	4
Vorauswahl	2					
Fachkenntnisse		1	**0**	**0**	**0**	**0**
Erfahrung		1	**0**	**0**	**0**	**0**
Methodenkompetenzen	6					
Analyseverhalten		2	**0**	**1**	**2**	**1**
Inhalte zusammenfassen			0	1	1	0
Bedarf erkennen			0	2	1	0
Tabellen interpretieren			0	1	1	0
Unternehmerisches Denken, Innovation		2	**0**	**1**	**2**	**0**
Präsentation		1	**0**	**0**	**1**	**0**
Organisation, Planung		1	**0**	**0**	**1**	**0**
Soziale Kompetenzen	6					
Führungsverhalten		2	**2**	**0**	**0**	**0**
Motivation			2	0	0	0
Delegation			1	0	0	0
Steuerung			1	0	0	0
Kommunikationsverhalten		1	**1**	**1**	**0**	**0**
Durchsetzungsvermögen		1	**1**	**1**	**0**	**0**
Kundenwirksamkeit		2	**1**	**2**	**0**	**0**
Persönliche Kompetenzen	3					
Leistungsmotivation		1	**0**	**1**	**1**	**0**
Belastbarkeit		1	**1**	**0**	**1**	**0**
Gewissenhaftigkeit		1	**0**	**1**	**1**	**0**

sind, gleich eins zu setzen, anderenfalls gleich null. Das Gewicht einer Merkmalsgruppe sollte in diesem Fall gleich der Anzahl der Merkmale der Gruppe gesetzt werden (Beispiel siehe Tabelle 4.2, Gruppe „Persönliche Kompetenzen“).

Bei der Wahl der Gewichte ist auf Ausgewogenheit besonders zu achten, damit nicht einzelne Merkmale übermäßig dominieren. Eine ausgewogenen Vergabe der Gewichte wird sehr erleichtert, wenn je nach Fragestellung jeweils einige ähnliche oder verwandte Merkmale in sinnvolle Gruppen zusammengefasst oder Merkmale operationalisiert werden, d.h. in leichter zu beobachtende Merkmale oder Verhaltensanker aufgeteilt werden. Dann brauchen jeweils nur wenige Gruppen, Merkmale oder Verhaltensanker hinsichtlich der Gewichte miteinander verglichen zu werden. Das ist ein besonderer Vorteil der Gruppierung. Die Merkmale sollten genügend unabhängig definiert sein. Das ist manchmal nicht der Fall, wenn sie zwar unterschiedlich benannt werden, aber doch eine ähnliche oder entgegengesetzte Bedeutung haben. Durch eine geeignete Vergabe z.B. kleinerer Gewichte im Verhältnis zueinander lässt sich erreichen, dass die Merkmale in einem solchen Fall beim Zusammensetzen nicht dominieren.

Wenn das aktuelle AC schon auf viele Teilnehmer angewendet und nach dem hier entwickelten Verfahren ausgewertet worden ist, liegen Unsicherheitskennwerte der Übungen bezüglich der Merkmale als Erfahrungswerte vor (Abschnitt 6.2.2). Dann lassen sich die Gewichte der Übungen bezüglich der Merkmale (in Tabelle 4.1 mit $+$ bezeichnet) auch so festlegen: Ist u ein solcher Unsicherheitskennwert, so ist das entsprechende Gewicht gleich $1/u^2$ anzusetzen. Eine Übung mit kleinem Unsicherheitskennwert bezüglich eines Merkmals im Vergleich zu anderen erhält auf diese Weise größeres Gewicht aufgrund der Erfahrungen aus vorangegangenen Anwendungen desselben AC. Der Ansatz folgt aus der Anwendung der *Methode der kleinsten Quadrate* bei der analogen Aufgabe beim physikalischen Messen, nämlich der Bildung eines Gesamtmittelwertes aus mehreren Messreihen bei Vergleichsmessungen (DIN 1319-4, 1999), wobei die Messreihen den Übungen hier entsprechen. Der Ansatz lässt sich nur bei den Gewichten der Übungen bezüglich der Merkmale heranziehen. Denn diese Gewichte sind haupsächlich dadurch bestimmt, wie genau ein betrachtetes Merkmal in einer Übung ermittelt werden kann, während alle anderen Gewichte im Wesentlichen davon abhängen, wie bedeutsam oder wichtig eine Merkmalsgruppe, ein Merkmal oder ein Verhaltensanker für das AC ist.

Nachdem Werte für alle Gewichte in die Matrix der Merkmale und Übungen nach Tabelle 4.1 eingetragen sind, muss klar sein, wie sie in das Auswerteverfahren einzubringen sind. Wenn Merkmale zum Gesamtmerkmal einer Merkmalsgruppe oder wenn

die Gesamtmerkmale der Merkmalsgruppen zum Teilnehmer-Gesamtmerkmal zusammengesetzt werden, sind die Gewichte, die in Tabelle 4.1 an den mit $\circ$ bzw. $\bullet$ bezeichneten Stellen stehen, als die Werte g_l in den Gleichungen (25) und (26) (die den Gleichungen (41) bzw. (42) entsprechen) zu verwenden. Die Gewichte an den mit $+$ bezeichneten Stellen der Matrix sind die Gewichte $P(A)$ in den Gleichungen (13) und (27) (die den Gleichungen (39) bzw. (40) entsprechen), abgesehen von der Normierung, die das Auswerteprogramm durchführen sollte. Dies gilt nur für Merkmale, bei denen Verhaltensanker nicht zu bewerten sind. Anderenfalls muss das Gewicht an der mit $+$ bezeichneten Stelle mit dem Gewicht des Verhaltensankers an der darunterstehenden mit $\times$ bezeichneten Stelle multipliziert werden, um das Gewicht $P(A)$ für die Aussagen A zu diesem Verhaltensanker in der betrachteten Übung zu erhalten. Tabelle 4.2 zeigt ein konkretes Beispiel einer Matrix der Merkmale und Übungen mit Gewichten zu einem AC. Diese Matrix ist nach dem Schema in Tabelle 4.1 aufgebaut.

Im Prinzip müsste jeder einzelnen Aussage A ein eigenes Gewicht $P(A)$ zugewiesen werden, z.B. auch hinsichtlich des Beobachters, von dem sie stammt. Die Unsicherheitskennwerte der Beobachter (Abschnitt 6.2.4) könnten herangezogen werden, um die Gewichte dazu, wie im vorangehenden Absatz beschrieben, festzulegen. Dies wird jedoch nicht weiter verfolgt, um das Auswerteverfahren für die Praxis nicht zu kompliziert zu gestalten. Mehreren vorliegenden Aussagen zu einem Merkmal oder Verhaltensanker in einer bestimmten Übung, die von mehreren Beobachtern stammen können, werden daher implizit wie üblich generell gleiche Gewichte zugewiesen, was in fast allen Fällen akzeptabel sein dürfte.

Die Gewichte stellen nichtstatistische Information dar, sie verhalten sich nicht zufällig. Ihre Unsicherheit kann daher nur mit Hilfe der Bayes'schen Statistik berücksichtigt werden. Soll dies geschehen, muss auch diese Unsicherheit angesetzt werden. Für die Gleichungen (40) und (42) (die den Gleichungen (27) bzw. (26) entsprechen) werden die globalen relativen Standardunsicherheiten u_r und u'_r nach Abschnitt 3.4.4 benötigt. Realistisch erscheint es, dafür etwa 10 % bis 20 % anzusetzen, also Werte von 0,1 bis 0,2. Es kann $u'_\mathsf{r} = u_\mathsf{r}$ gesetzt werden.

Abschließend wird empfohlen, für jedes, auch zusammengesetztes Merkmal, wenn sinnvoll, einen *Akzeptanzgrenzwert* auf der Bewertungskala festzulegen, womit das Auswerteprogramm eine Warnmeldung erzeugen kann, wenn der sich für das Merkmal ergebende Unsicherheitsbereich nicht vollständig auf der besseren Seite des Akzeptanzgrenzwertes liegt. (Abschnitt 4.5.1 und Anhang B, Anlagen 1 und 2)

4.3.2 Wahl der Bewertungsskala

In Abschnitt 3.4.1 wurde bereits beschrieben, wie eine gestufte Bewertungsskala gebildet wird. Hier geht es darum, eine geeignete Bewertungsskala für ein aktuelles AC auszuwählen.

Alle im AC gebräuchlichen Bewertungsskalen sind gestuft, was auch zweckmäßig ist. Steigende oder fallende Skalen, bei denen die obere bzw. die untere Skalengrenze die beste mögliche Bewertung bedeutet, und solche mit einer geraden oder ungeraden Anzahl N der Stufen kommen vor, z.B. bei den AC-Serien, die hier für die Evaluierung verwendet werden (Abschnitt 5.2). Eine gerade Anzahl der Stufen wird manchmal gewählt, um zu verhindern, dass weniger entscheidungsfreudige Beobachterpersonen zu oft auf der mittleren Stufe bewerten. Meistens sind die Skalen von 1 bis N mit der Stufenhöhe 1 gestuft. Eine Stufenhöhe 0,5 (Abschnitt 5.2) oder eine untere Skalengrenze ungleich 1 kommen aber auch vor. Skalenbeispiele aus der Praxis sind (Stufenhöhe in Klammern): 1 bis 4 (1), 1 bis 5 (0,5), 1 bis 6 (1), 70 bis 130 (1). Bei der letztgenannten Skala werden die Bewertungen in Prozent angegeben, die Skalenmitte liegt bei 100 % (siehe Beispiel in Anhang B, Anlagen 1 und 2). Es gibt auch verbal definierte Bewertungsskalen, z.B. mit den Stufen „sehr gut“, „gut“, „befriedigend“, „ausreichend“, „mangelhaft“ oder „niedrig“, „mittel“, „hoch“.

Für die Auswertung ist es sehr zweckmäßig, nur eine einzige Bewertungsskala für alle Bewertungen im AC vorzusehen. Das vereinfacht Berechnungen und graphische Darstellungen der Ergebnisse erheblich. Nun wird bei Tests aber oft die Bewertung eines Merkmals als erzielte Punktzahl angegeben, d.h. es wird auf einer Punktskala von 0 bis M bewertet, wobei M die maximal erzielbare Punktzahl ist. Meist ist M viel größer als die Anzahl N der Stufen der gewählten Bewertungsskala. In diesem Fall ist die Punktskala auf die Bewertungsskala zu normieren, d.h. es ist jede mögliche Punktzahl einer Stufe der Bewertungsskala und jede größere Punktzahl derselben oder einer nachfolgenden Stufe so zuzuweisen, dass jede Stufe nach Möglichkeit gleich viele mögliche Punktzahlen umfasst.

Für das Auswerteverfahren dieser Arbeit sind nur nummerische, gestufte Bewertungsskalen geeignet. Empfohlen wird eine gestufte Skala von 1 bis N oder von 0 bis $N-1$ mit der Stufenhöhe 1. Die Stufenhöhe 1 ist nicht unbedingt erforderlich, aber zweckmäßig. Anderenfalls muss dafür gesorgt werden, dass Zwischenstufen von den Beobachtern nicht vernachlässigt werden. Denn die Bewertungsskala sollte so beschaffen sein, dass keine Stufe sich von vornherein auf irgendeine Weise von den anderen unterscheidet. Zweckmäßig ist es, eine Bewertungsskala mit nicht zu vielen, aber auch

nicht zu wenigen Stufen zu wählen. Bei einer Skala mit vielen Stufen, z.B. von 1 bis 100, ist es für den Beobachter schwierig zu entscheiden, ob er z.B. mit 86 oder 87 bewerten soll, andererseits liefert eine Skala mit nur wenigen Stufen einen hohen Beitrag zur Unsicherheit. Denn die Skalenstufung erzeugt Unsicherheit über die zu bewertenden Merkmale, je weniger Stufen, desto mehr. Im trivialen, nicht sinnvollen Extremfall liefert eine Skala mit nur einer einzigen Stufe gar keine vernünftige Information, weil sie dem Beobachter keine Freiheit für die Bewertung bietet. Bei sehr vielen Stufen ist dagegen für den Beobachter die Festlegung auf eine Bewertung schwierig. Ein praktikables Optimum scheint bei etwa $N = 8$ bis höchstens etwa 15 Stufen zu liegen. Danach ist eine Bewertungsskala mit etwa 8 bis 15 Stufen zu empfehlen. Dabei bleibt die durch die Stufung erzeugte Unsicherheit ausreichend klein. Damit ist es für den Beobachter noch nicht allzu zu schwierig, seine Bewertung festzulegen, und er erhält genügend Freiheit, seine eigene Unsicherheit durch Angabe eines Bewertungsbereichs auszudrücken, d.h. er hat genügend viele Möglichkeiten, die minimalen und maximalen Bewertungen zu wählen und diese auch z.B. einfach durch Markieren, Ankreuzen oder Anklicken von Zahlen 1 bis N oder 0 bis $N-1$ auf einem dementsprechend geeignet gestalteten Formularblatt oder in eine Bildschirmmaske abzugeben (Abschnitt 4.3.4, Beispiele siehe Tabellen 4.3 und 4.4).

4.3.3 Bewertung

Siehe hierzu auch die Abschnitte 1.4, 2.2.5 und 3.4.1.

Die Beobachter werden angewiesen, in einer Übung zu jedem zu bewertenden Merkmal X oder Verhaltensanker eines Teilnehmers eine Aussage A abzugeben, die aus einer minimalen Bewertung und einer maximalen Bewertung besteht. Diese Bewertungen können auch gleich sein. Sie kennzeichnen die unterste bzw. oberste Stufe (Teilintervall) der Bewertungsskala, auf die der Beobachter hinsichtlich der Ausprägung des Merkmals X oder Verhaltensankers beim Teilnehmer aufgrund der ihm aus der Übung zugeflossenen Information nach eigenem Dafürhalten setzen würde (Abschnitt 3.4.1). Besteht die Aussage aus nur einer Bewertung, so bedeutet dies ebenfalls, dass minimale und maximale Bewertung mit jener übereinstimmen. Fehlen zu einer vorgesehenen Aussage beide Bewertungen, was nicht selten vorkommt, z.B. wenn der Beobachter meint, aus den Übung keinerlei Information gewonnen zu haben, so bedeutet dies, dass die minimale und die maximale Bewertung mit der kleinsten bzw. der größten möglichen Bewertung auf der Skala, d.h. mit der unteren bzw. oberen Skalengrenze übereinstimmt.

Warum ist es nun in der Praxis zweckmäßig, dass der Beobachter seine eigene Bewertungsunsicherheit durch Angabe einer minimalen und maximalen gerade noch sinnvoll

erscheinenden Bewertung ausdrückt? Es gibt auch Alternativen [Uns] (Jungermann, Pfister und Fischer, 1998) das nötige Mehr an Information zur Quantifizierung der Unsicherheit zu gewinnen, z.B. durch „lautes Denken“ (Mattenklott, 1988), jedoch erfordern diese vom Beobachter zusätzliche Überlegungen, Berechnungen oder Schreibarbeit, die ihm im laufenden AC kaum zugemutet werden kann, oder sie sind nur in Sonderfällen anwendbar. Deshalb scheint die hier gewählte Art der Bewertung die für die Praxis einfachste Möglichkeit darzustellen. Außerdem besteht als weiterer Vorteil nach Abschnitt 3.2.2 bei gemeinsamer Bewertung zweier Merkmale oder Verhaltensanker keine Korrelation. Auch beim Messen in der Physik ist es üblich, die Unsicherheit einer Einflussgröße durch realistische Abschätzung eines kleinsten und eines größten möglichen Wertes zu ermitteln [Uns]. Schließlich ist noch der Vorteil zu erwähnen, der darin liegt, dass die Wahrscheinlichkeitsverteilung zu einem Merkmal nicht explizit aufgestellt zu werden braucht. Erwartungswert und Varianz können aus der minimalen und maximalen Bewertung direkt berechnet werden.

Eine Alternative, die den Mehraufwand deutlich werden lässt, sei hier als Beispiel betrachtet. Möglich wäre es, dass der Beobachter aufgrund der in der Übung gewonnenen Information eine mittlere Bewertung als besten Schätzwert x eines Merkmals X sowie direkt die zugehörige Standardunsicherheit $u(x)$ ermittelt und angibt. Aber wie kann der Beobachter diese Werte ermitteln? Er muss sich wie auch auf dem empfohlenen Weg Gedanken machen über eine minimal und eine maximal sinnvolle Bewertung, die mit $x - u(x)$ bzw. $x + u(x)$ bezeichnet seien, aber zusätzlich deren Summe und Differenz bilden und halbieren, um x und $u(x)$ zu erhalten. Diese muss er dann mit Kommastellen niederschreiben und in den Computer eingeben. Eine umständliche Prozedur. Die Berechnungen kann der Computer erledigen. Deshalb genügt es, die minimal und maximal sinnvollen Bewertungen anzugeben.

Wenn allerdings x und $u(x)$ bereits gegeben sind, liegt es natürlich nahe, sie direkt zu verwenden. Dieser Fall liegt beispielsweise vor, wenn eine Vorauswahl der Teilnehmer in Form eines (einfachen) AC durchgeführt worden ist und ein Merkmal der Vorauswahl, z.B. deren Gesamtmerkmal, als Merkmal X des aktuellen AC vorgesehen wird, um das Vorauswahlergebnis x dieses Merkmals und die zugehörige Standardunsicherheit $u(x)$ zu berücksichtigen.

Es kann auch sein, dass die Beobachter n i c h t angewiesen werden, wie oben beschrieben zu verfahren (z.B. bei der AC-Serie ST, Abschnitt 5.4.3). Der Beobachter wird dann an die Unsicherheit keinen Gedanken verschwenden und einfach eine Bewertung angeben, die er als sinnvoll erachtet. Wenn mehrere Beobachter dasselbe Merkmal bewerten oder mehrere Verhaltensanker des Merkmals zu bewerten sind, ist

das nicht weiter schlimm, weil dann genügend Information zur Quantifizierung der Unsicherheit des Merkmals, u.a. aus der Streuung der Werte, verfügbar ist. Wenn jedoch der Beobachter allein das Merkmal bewertet, wird es schwieriger, $u(x)$ zu ermitteln. Bei einem Test, der nach Abschnitt 2.2.3 sowohl als Übung als auch als Beobachter aufgefasst wird, ist das beispielsweise so. Meist liegt nur das Testergebnis als erzielte Punktzahl vor und sonst nichts. Bei Anwendung eines gut evaluierten Tests ist aber davon auszugehen, dass die zum Testergebnis gehörende Standardunsicherheit nur einige wenige Punkte beträgt. Wenn dann die maximale Anzahl M der erzielbaren Punkte wesentlich größer ist als die Anzahl N der Stufen der Bewertungsskala, darf angenommen werden, dass die Unsicherheit zum Testergebnis wesentlich kleiner ist als die in Betracht zu ziehende Unsicherheit aufgrund der Skalenstufung und daher vernachlässigt werden kann. Dann sind nach Zuweisung der erzielten Punkzahl auf eine Stufe der Bewertungsskala die so festgelegte Bewertung als mit der minimalen und der maximalen Bewertung übereinstimmend zu betrachten.

Wenn das AC bereits auf sehr viele Teilnehmer angewendet worden ist, gibt es eine weitere Möglichkeit, $u(x)$ festzulegen. Dann lässt sich aus den schon angefallenen Daten eine „mittlere“ Standardunsicherheit berechnen, den Unsicherheitskennwert u_M des Merkmals (Abschnitt 6.2.2), und als $u(x)$ benutzen. Das hat allerdings einen Nachteil: die Unsicherheiten des Merkmals sind dann bei allen Teilnehmern gleich groß. Die beiden Ergebnisse zu zwei Teilnehmern lassen sich dann zwar noch nach Abschnitt 3.3.3 kritisch miteinander vergleichen, wenn sie sich aber nicht signifikannt unterscheiden, ist es nicht mehr möglich, eine Entscheidung für den Teilnehmer mit der kleineren Unsicherheit zu treffen. Das dürfte in der Praxis aber nur eine geringe Rolle spielen, weil meist noch genügend viele andere Merkmale in Betracht gezogen werden können und müssen.

Auf einen möglichen Einwand gegen die vorgeschlagene Art zu bewerten sollte noch eingegangen werden. Er lautet: Wenn eine einzelne Bewertung nur unsicher angebbar ist, dann sind es deren zwei, die minimal und die maximal sinnvolle, natürlich auch. Das sollte für beide je weitere minimale und maximale Bewertungen erfordern usw., also schließlich einen uferlosen Aufwand. Nach Weise und Wöger (1999) bringt hier jedoch höherer Aufwand praktisch keinen Gewinn. Auch werden nach der Bayes'schen Statistik alle Schlüsse nur anhand der gerade vorliegenden Information gewonnen. Zwei Bewertungen bieten mehr Information als eine Bewertung. Dieses Mehr wird zur Quantifizierung der Unsicherheit benutzt. Noch mehr Information könnte die Situation wohl im Prinzip verbessern, doch wäre der Aufwand, diese Information zu gewinnen, unvertretbar. Das gilt nicht nur hier, sondern auch bei sehr genauen Messungen in der Physik.

4.3.4 Erfassung der Bewertungen

Die sofortige Auswertung der im laufenden AC anfallenden Bewertungen und anderen Daten mittels EDV erfordert eine wohlorganisierte Datenerfassung. Es fallen sehr viele Daten an, die schnell und fehlerfrei in den Computer eingegeben werden müssen, damit die Ergebnisse der Auswertung bereits in der Beobachterkonferenz als Beratungsunterlage zur Verfügung stehen. Während bei der Auswertung die Berechnungen und die Ergebnisausgabe in Sekundenschnelle automatisch ablaufen können, ist die Dateneingabe mühselig und fehleranfällig und kann deshalb in der Praxis das Hauptproblem für die rechtzeitige Auswertung bilden, wenn sie nicht sorgfältig vorbereitet und gut organisiert wird. Dies zeigte schon die nachträgliche Eingabe der Daten der drei AC-Serien, die für die Evaluierung des Auswerteverfahrens herangezogen wurden (Kapitel 5 und 6). Aufgrund der hierbei gesammelten Erfahrungen sollte die Datenerfassung wie folgt verlaufen:

Im AC ist es üblich und sinnvoll, dass jeder Beobachter zu jedem Teilnehmer und zu jeder Übung ein Formularblatt erhält. Darauf kann er während der Übung Notizen zum Verhalten des Teilnehmers schreiben, die ihm für die Bewertung wichtig erscheinen. Außerdem vermerkt der Beobachter auf dem Blatt nach der Übung seine Bewertungen zu jedem in dieser Übung zu bewertenden Merkmal und Verhaltensanker. Die so ausgefüllten Formularblätter, insgesamt oft mehrere hundert Stück, dienen dann als Vorlage für die Dateneingabe und als Beratungsunterlage für die Beobachterkonferenz.

Für die Auswertung aller Bewertungen mittels EDV sollte für jede einzelne Übung ein spezielles Bewertungsformular zweckmäßig gestaltet werden. Es sollte Felder aufweisen, in die Kenndaten und die Bewertungen vom Beobachter sorgfältig einzutragen sind. Dies erleichtert die Dateneingabe in den Computer, die möglichst sofort nach jeder Übung entweder vom Beobachter selbst oder von einer Hilfskraft vorgenommen werden sollte. Zweckmäßig ist es, die Daten mittels der Tastatur in vorbereitete Felder einer Maske auf dem Bildschirm einzugeben, die dem jeweiligen Bewertungsformular entspricht, oder schneller mittels eines Scanners. Für diesen Zweck lassen sich z.B. das bekannte Programm Excel von der Firma Microsoft bzw. Formular-Leseprogramme, wie TELEform von der Firma Electric Paper, verwenden. Ein Beispiel für die mögliche Gestaltung eines Bewertungsformulars zeigt Tabelle 4.3. Hinsichtlich der Dateneingabe kommt es dabei nur auf die zweckmäßige Anordnung der Kenndaten- und Bewertungsfelder an. Ansonsten kann das Formular nach Bedarf im AC beliebig gestaltet werden.

Das in Tabelle 4.3 als Beispiel vom Papierformat DIN A4 auf DIN A5 verkleinert dargestellte Bewertungsformular besteht aus einem Kopf für Kenndaten und einem

Rumpf für die in der Übung zu bewertenden Merkmale und Verhaltensanker. Der Kopf umfasst eine Titelzeile, die Blattnummer sowie Name und Nummer des Teilnehmers, des Beobachters und der Übung. Die Ziffern der Nummern sind in die dafür vorgesehenen rechteckigen Felder einzutragen. Die Blattnummer dient neben den anderen Nummern der Identifizierung. Sie kann für computerinterne Zwecke oder für die Archivierung erforderlich oder sinnvoll sein, z.B. dazu dienen, auf den Datensatz des Blattes nachträglich schnell zuzugreifen, um ihn zu präsentieren oder zu korrigieren, oder um den Datenbestand auf Vollständigkeit zu prüfen. Der Kopf des Formulars kann schon vor der Übung ausgefüllt werden, z.B. von einer Hilfskraft oder auch durch den Computer selbst, wenn alle Blätter vor der Verteilung an die Beobachter ausgedruckt werden.

Der Rumpf des Formulars enthält zu jedem zu bewertenden Merkmal und Verhaltensanker zwei Felder für die minimale und maximale Bewertung und daneben Platz für kurze Notizen. Nötigenfalls können für letztere auch der freie Raum unterhalb des Rumpfes und die Rückseite des Formulars benutzt werden. Für die Bewertungen ist im Beispiel eine zehnstufige Bewertungsskala von 0 bis 9 zugrunde gelegt. Wenn minimale und maximale Bewertung übereinstimmen, braucht nur eines der Felder ausgefüllt zu werden (Zeilen 5 und 6). Bei Enthaltung der Bewertung, z.B. weil das Merkmal nicht beobachtet werden konnte, sind die Felder leer zu lassen (Zeile 9) oder die minimal und maximal möglichen Bewertungen, d.h. die unteren und oberen Skalengrenzen in die Felder einzutragen (Zeile 10). Bei fehlerhafter Eintragung ist diese durchzustreichen (Zeile 11) und für die Berichtigung eine Korrekturzeile mit Hinweis auf die fehlerhafte Zeile zu verwenden (Zeile 12). Die im Formularbeispiel aufgeführten Merkmale und Verhaltensanker dienen haupsächlich der Verdeutlichung, sie sind aber mit Bezug zu einem realen AC in der Praxis gewählt. Das Beispiel gilt für die Übung 1 „Mitarbeitergespräch“ des in Tabelle 4.2 dargestellten AC.

Tabelle 4.3: Beispiel eines Bewertungsformulars

AC Einstellung Filialleiter Juli 2002		Nr.	Zeile
	Blatt:	0157	1
Teilnehmer:	Meyer	19	2
Beobachter:	Brinkmann	05	3
Übung:	Mitarbeitergespräch	02	4
Merkmal Verhaltensanker		Bewertung von – bis	Zeile
Führungsverhalten			
Motivation		3	5
Delegation		04	6
Steuerung		79	7
Kommunikationsverhalten		27	8
Durchsetzungsvermögen			9
Kundenwirksamkeit		09	10
Belastbarkeit		~~97~~	11
	Korrektur zu Zeile: 11	56	12
	Korrektur zu Zeile:		13

Tabelle 4.4: Drei Beispiele für die Angabe von Nummern und Bewertungen im Bewertungsformular

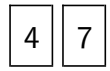
1) Eintragen von Ziffern in Kästchen	
2) Auffüllen von Lücken in einem Strichcode	
3) Markieren von Ziffern auf einer Skala	

Tabelle 4.4 zeigt drei Beispiele, wie die Angabe von Nummern und Bewertungen im Bewertungsformular vorgesehen werden kann. Beispiel 1, das Eintragen von Ziffern in Kästchen, ist bei dem in Tabelle 4.3 dargestellten Formular verwendet. Alternativ dazu sind im Beispiel 2 Nummern und Bewertungen als Strichcode anzugeben, der mit Hilfe eines Handscanners wie an einer Supermarkt-Kasse gelesen werden kann. Dabei stehen auf dem Formular zunächst nur Folgen von je 11 Strichen. Der Beobachter füllt z.B. mit einem schwarzen Filzstift die zu seinen Bewertungen gehörenden Lücken zwischen den Strichen aus. Beispiel 3 besteht darin, Ziffern von Nummern und Bewertungen auf einer Skala zu markieren. Jedes dieser Beispiele hat Vor- und Nachteile. Beispiel 1 erfordert auf dem Formular weniger von dem für Notizen benötigten, aber karg bemessenen Platz als die beiden anderen Beispiele. Dagegen sind bei diesen die Nummern und Bewertungen leichter maschinell lesbar als bei Beispiel 1. Nachteilig bei dem Strichcode in Beispiel 2 ist, dass die Lücken sehr sorfältig ausgefüllt werden müssen, was sich im laufenden AC nur schwer durchsetzen lässt. Bei den Beispielen 2 und 3 ist für die Nummern im Formularkopf noch eine Besonderheit zu beachten: Strichcode und Ziffernskala lassen nur Nummern mit nach rechts größer werdenden Ziffern zu, weil z.B. 21 nicht von 12 unterschieden wird und 22 nicht eingegeben werden kann, wodurch 21 und 22 als mögliche Nummern ausfallen. Die dadurch entstehenden Lücken in der Nummerierung bedeuten jedoch nur eine geringe Einschränkung, weil der Strichcode und die Ziffernskala, wie sie in Tabelle 4.4 gezeigt sind, eine sicher ausreichende Anzahl von $2^{10} = 1024$ unterschiedlichen Nummern zulassen, da jede der 10 Ziffern entweder markiert wird oder nicht.

4.4 Berechnung der Teilnehmer-Ergebnisse und zugehörigen Unsicherheiten

Das Auswerteverfahren wird nun in den folgenden Schritten durchgeführt (siehe auch Abschnitte 3.4.1, 3.4.3, 3.4.4, 4.2 und 6.1):

Der erste Schritt besteht darin, zu jedem Merkmal X eines Teilnehmers alle Aussagen A der Beobachter aus allen Übungen zu sammeln. Zu den Aussagen siehe die Abschnitte 3.4.1 und 4.3.3. Wenn ein Merkmal X durch Verhaltensanker operationalisiert ist und nicht das Merkmal selbst, sondern diese Verhaltensanker einzeln bewertet sind, so werden zweckmäßig alle Aussagen zu den Verhaltensankern als Aussagen zum Merkmal X gezählt. Alle gesammelten Aussagen zum Merkmal X des Teilnehmers werden von $A = 1$ bis M nummeriert, wobei M die Anzahl dieser Aussagen ist.

Der zweite Schritt umfasst die folgenden Berechnungen: Für jede Aussage A werden die untere Intervallgrenze t_{j-1} des Teilintervalls zur minimalen Bewertung j der Aussage und die obere Intervallgrenze t_k des Teilintervalls zur maximalen Bewertung k der Aussage gebildet ($j \le k$). Ist die Bewertungsskala z.B. mit der Stufenhöhe 1 ganzzahlig gestuft, so sind $t_{j-1} = j - 0{,}5$ und $t_k = k + 0{,}5$. Schließlich sind noch Gewichte $P(A)$ für die Aussagen A heranzuziehen, z.B. $P(A) = 1/M$, wenn keine der Aussagen bevorzugt werden kann oder soll, außerdem die globale relative Standardunsicherheit u'_{r} dieser Gewichte (Abschnitte 3.4.1 und 4.3.1). Nach den Gleichungen (11), (13) und (27) werden dann nacheinander berechnet:

$$\mathrm{E}\,X|A = (t_k + t_{j-1})/2 \qquad \mathrm{E}\,X^2|A = (t_k^2 + t_k t_{j-1} + t_{j-1}^2)/3 \tag{38}$$

$$x = \mathrm{E}\,X = \sum_{A=1}^{M} (\mathrm{E}\,X|A) P(A) \qquad \mathrm{E}\,X^2 = \sum_{A=1}^{M} (\mathrm{E}\,X^2|A) P(A) \tag{39}$$

$$u^2(x) = \mathrm{E}\,X^2 - x^2 + \sum_{A=1}^{M} P^2(A)\,(\mathrm{E}\,X|A - x)^2 u_{\mathrm{r}}'^2 \tag{40}$$

Damit sind der beste Schätzwert x und die zugehörige Standardunsicherheit $u(x)$ des Merkmals X des Teilnehmers gewonnen. Bei allen Merkmalen wird so verfahren.

Im dritten Schritt werden die im zweiten Schritt nach Gleichung (39) berechneten Schätzwerte x_l aller L Merkmale X_l von Gleichung (37) sowie die zugehörigen Standardunsicherheiten $u(x_l)$ nach Gleichung (40) in die Gleichungen (25) und (26) eingesetzt. Dadurch ergeben sich schließlich der Schätzwert z und die zugehörige

Standardunsicherheit $u(z)$ des zusammengesetzten Merkmals Z eines Teilnehmers wie folgt:

$$z = \sum_{l=1}^{L} h_l \, x_l \qquad \left(h_l = \frac{g_l}{\sum_{j=1}^{L} g_j} \right) \tag{41}$$

$$u^2(z) = \sum_{l=1}^{L} h_l^2 \left(u^2(x_l) + (x_l - z)^2 u_{\mathsf{r}}^2 \right) \tag{42}$$

Die Schätzwerte h_l für die normierten Gewichte H_l sind hierbei aus den Gewichtswerten g_l nach Tabelle 4.1 zu bilden, außerdem ist die globale relative Standardunsicherheit u_{r} zu den Gewichten heranzuziehen (Abschnitt 4.3.1).

Wichtige Voraussetzung für Gleichung (42) ist, dass die Merkmale X_l eines Teilnehmers nicht korreliert sind (Abschnitt 3.5.2). Dies wird in Abschnitt 6.3.3 evaluiert. Es erweist sich dort, dass diese Voraussetzung als in ausreichender Näherung erfüllt angesehen werden kann.

Außerdem ist es im vierten Schritt zweckmäßig, für jedes Merkmal eine Rangfolge der Teilnehmer nach den ermittelten Schätzwerten des Merkmals zu bilden und anhand des vorgegebenen Akzeptanzgrenzwertes des Merkmals (Abschnitt 4.3.1) festzustellen, welche Schätzwerte der Teilnehmer als ausreichend akzeptiert werden können und welche nicht. Ein Akzeptanzkriterium für diesen Zweck wird in Abschnitt 4.5.1 vorgeschlagen.

4.5 Beurteilung der Unsicherheit

Sehr wichtig ist es, die nach Abschnitt 4.4 berechneten Standardunsicherheiten zu den Teilnehmer-Ergebnissen richtig zu verstehen und zu beurteilen. Darum werden zunächst in Abschnitt 4.5.1 die wesentlichen Ideen zur Auffassung der Unsicherheit auf dem Boden der Bayes'schen Staatistik aus den Abschnitten 3.3.1, 3.3.2 und 3.4.4 noch einmal kurz herausgearbeitet. Daraus folgt auch das am Ende von Abschnitt 4.4 erwähnte Akzeptanzkriterium. In Abschnitt 4.5.2 wird dann beschrieben, wie die berechneten Standardunsicherheiten in ihrer nummerischen Größe zu beurteilen sind.

4.5.1 Interpretation der Unsicherheit, Akzeptanzkriterium

Das nach Abschnitt 4.4 berechnete Ergebnis x oder z ist der beste Schätzwert eines einfachen Merkmals X bzw. eines zusammengesetzten Merkmals Z eines Teilnehmers, z.B. eines Gruppenmerkmals oder des Gesamtmerkmals des Teilnehmers. Das Ergebnis

x oder z ergibt sich als der Erwartungswert und die zugehörige Standardunsicherheit $u(x)$ bzw. $u(z)$ als die Standardabweichung einer Wahrscheinlichkeitsverteilung, die nichts als die aktuelle Kenntnis des Merkmals aufgrund der gerade vorliegenden unvollständigen Information wiedergibt. Diese Verteilung ist k e i n e Verteilung von Werten, wie sie bei wiederholten Versuchen zufällig auftreten, sondern basiert auf dem Bernoulli'schen Prinzip, gleichartigen Möglichkeiten die gleiche Wahrscheinlichkeit zuzuweisen, wie die Wahrscheinlichkeit 1/2 für jede Seite einer Münze, o h n e dass diese jedoch wirklich geworfen wird (Abschnitt 3.1).

Die Unsicherheit des Merkmals eines Teilnehmers bedeutet also n i c h t , dass das Merkmal in seiner Ausprägung schwankt oder dass diese Ausprägung sich tatsächlich im Unsicherheitsbereich befindet, sondern nur, dass das Merkmal nicht genau bekannt ist. Die Unsicherheit ist dementsprechend ein Maß für den Mangel in der Kenntnis des Merkmals, ein Maß für den Mangel an Information darüber und in dieser Hinsicht auch ein Maß für die Qualität des Teilnehmer-Ergebnisses. Das heißt aber auch nicht, dass dieses Ergebnis schlechter ist als ein gleiches Ergebnis eines anderen Teilnehmers, zu dem aber eine kleinere Unsicherheit gehört, sondern z.B., dass wegen des Informationsmangels ein höheres Entscheidungsrisiko besteht. Gründe für erheblichen Informationsmangel können sein, dass Übungen für manche Merkmale nicht aussagekräftig genug sind, dass der Teilnehmer unklares und daher schlecht bewertbares Verhalten zeigt oder dass Beobachter nicht fähig genug sind, vorliegende Information richtig zu bewerten. Die Unsicherheit zu einem berechneten Ergebnis umfasst nach Abschnitt 3.4.4 sowohl die Unsicherheiten zu den Einzelbewertungen der beitragenden Merkmale der Teilnehmer durch die Beobachter, die durch die Angabe jeweils minimaler und maximaler Bewertungen ausgedrückt werden, als auch die Streuung der Bewertungen durch die einzelnen Beobachter in den einzelnen Übungen, außerdem den Unsicherheitsbeitrag der Skalenstufung sowie global die Unsicherheiten der Gewichte.

Der Unsicherheitsbereich eines Merkmals hat die Grenzen $x - u(x)$ und $x + u(x)$ bzw. $z - u(z)$ und $z + u(z)$. Er ist kein Vertrauensintervall der konventionellen Statistik, sondern der Bereich derjenigen Werte, die aufgrund der vorliegenden Information dem Merkmal als vernünftige Schätzwerte für dessen zu ermittelnde Ausprägung beim Teilnehmer zugewiesen werden können [Uns]. Der Unsicherheitsbereich zu einem aus mehreren (mindestens drei) Merkmalen zusammengesetzten Merkmal Z kann allerdings als Näherung eines Vertrauensintervalls zur Wahrscheinlichkeit 0,68 ($\approx 2/3$) aufgefasst werden, weil die dann zugrunde liegende Wahrscheinlichkeitsverteilung oft näherungsweise eine Normalverteilung ist. (Zum Vertrauensintervall zu anderen Wahrscheinlichkeiten siehe Abschnitt 3.4.4) Die Ergebnisse für zwei Teilnehmer zu demselben Merkmal sind nur dann signifikant verschieden, wenn die beiden zugehörigen

Unsicherheitsbereiche sich nicht überlappen (Abschnitt 3.3.3).

Mit dem Unsicherheitsbereich lässt sich auch ein Akzeptanzkriterium begründen. Wenn der Unsicherheitsbereich vollständig auf der besseren Seite eines auf der Bewertungskala für das Merkmal vorgegebenen Akzeptanzgrenzwertes a liegt, ist jeder vernünftige Schätzwert des Merkmals besser als a. Dann kann dem Teilnehmer bezüglich des betrachteten Merkmals ein positives Votum zugesprochen werden, z.B. ein Einstellungsvotum zum Gesamtmerkmal des Teilnehmers in einem Auswahl-AC (Abschnitt 6.1.1). Das Akzeptanzkriterium lautet daher: Das Ergebnis x kann als ausreichend gut akzeptiert werden, wenn $x + u(x) < a$ bei fallender Bewertungsskala oder wenn $x - u(x) > a$ bei steigender Bewertungsskala. Das Kriterium gilt für z entsprechend.

4.5.2 Große und kleine Unsicherheit

Noch wesentlich wichtiger als die Prüfung auf Erfüllung des Akzeptanzkriteriums ist es, zu klären, ob überhaupt genügend Information vorhanden ist, um die Qualität eines Ergebnisses als ausreichend gut ausweisen zu können. Die Standardunsicherheit kann auch dabei helfen, weil sie ja ein Maß für den Informationsmangel darstellt. Allerdings kann es einem berechneten nummerischen Wert der Standardunsicherheit nicht ohne Weiteres angesehen werden, ob er einem „großen“ oder „kleinen“ Informationsmangel entspricht. Dies lässt sich erst entscheiden durch Vergleich des vorliegenden Wertes der Standardunsicherheit mit charakteristischen Werten, die für jedes Merkmal gesondert durch Betrachtung extremer Fälle zu ermitteln sind, was allgemein bei Prüfungen auf Plausibilität wie hier oft hilfreich ist.

Um diese *charakteristischen Standardunsicherheiten* eines Merkmals zu gewinnen, werden deshalb drei extreme Fälle vorliegender Information zu dem Merkmal untersucht:

1) Die Information ist ideal. Bei jeder Aussage, die zum Merkmal beiträgt, stimmen die minimale und die maximale Bewertung überein. Es gibt auch keine Streuung der Bewertungen, d.h. alle Aussagen sind gleich.
2) Es liegt gar keine Information vor. Bei jeder Aussage, die zum Merkmal beiträgt, stimmt die minimale Bewertung mit der kleinsten möglichen Bewertung überein, ebenso die maximale Bewertung mit der größten möglichen Bewertung.
3) Die Information ist extrem widersprüchlich. Bei jeder Aussage, die zum Merkmal beiträgt, stimmen zwar die minimale und die maximale Bewertung überein, diese aber auch je zur Hälfte mit der kleinsten und größten möglichen Bewertung.

Der Idealfall 1 liefert offenbar die kleinste mögliche Standardunsicherheit, die charakteristische Standardunsicherheit u_1. Dagegen liegt im Fall 2 ein für die Praxis völlig inakzeptabler Informationsmangel vor. Eine berechnete Standardunsicherheit sollte unter

allen Umständen kleiner sein als die Standardunsicherheit in diesem Fall, die charakteristische Standardunsicherheit u_2. Im Fall 3 der widersprüchlichen Information liegen die Bewertungen und damit die möglichen Werte des Schätzers noch weiter auseinander als im Fall 2, was die Varianz weiter vergrößert und damit eine noch größere Standardunsicherheit als im Fall 2 zur Folge hat. Fall 3 kann deshalb als erst recht nicht akzeptabel außer Acht gelassen werden. Eine Standardunsicherheit ist also niemals kleiner als u_1 im Idealfall 1, und es ist zu fordern, dass sie w e s e n t l i c h kleiner ist als u_2 im Fall 2 der Desinformation. Wenn diese *Informationsbedingung* nicht erfüllt ist, liegt eine zu große Unsicherheit aufgrund eines nicht mehr vertretbaren Informationsmangels vor. Dann sollte das auswertende Computer-Programm eine Warnmeldung erzeugen. Wie „wesentlich kleiner“ genauer zu verstehen ist, wird weiter unten besprochen.

Die charakteristischen Standardunsicherheiten u_1 und u_2 lassen sich für ein einfaches Merkmal X aus Gleichung (40) gewinnen. In beiden Fällen wirken alle Aussagen wie eine einzige, sodass $M = 1$ und $P(A) = 1$ sowie $x = \mathrm{E}\,X|A$, wodurch die Summe in Gleichung (40) entfällt. Übrig bleibt die Varianz einer einfachen Rechteckverteilung der Breite B im Fall 1 und NB im Fall 2. Dabei ist B die Stufenhöhe und N ist die Stufenzahl der Bewertungsskala. Die Varianz hat den Wert $B^2/12$ im Fall 1 und $(NB)^2/12$ im Fall 2. Daraus folgen die charakteristischen Standardunsicherheiten $u_1 = B/\sqrt{12}$ und $u_2 = NB/\sqrt{12}$, d.h. $u_2 = Nu_1$. Bei einer Bewertungsskala von 1 bis N ist $B = 1$.

Für ein zusammengesetztes Merkmal Z ergeben sich die charakteristischen Standardunsicherheiten aus Gleichung (42). Darin sind für $u(x_l)$ die eben berechneten Werte für u_1 und u_2 einzusetzen. Im Fall 2 sind alle x_l untereinander gleich und auch gleich z nach Gleichung (41), weil die Summe der h_l gleich 1 ist. Deshalb ist $x_l - z = 0$. Das erbringt

$$u_2 = NB\sqrt{\frac{1}{12}\sum_{l=1}^{L} h_l^2} \tag{43}$$

Im Fall 1 ist als u_1 das minimale $u(z)$ zu suchen. Obwohl im Allgemeinen $x_l \neq z$, ist $x_l = z$ nicht ausgeschlossen. Deshalb kann für das Minimum $x_l - z = 0$ gesetzt werden. Es entsteht zunächst $u_1 = u_2/N$ mit u_2 nach Gleichung (43). Darin hat die Summe der h_l^2 ihren minimalen Wert $1/L$ bei gleichen Gewichten $h_l = 1/L$. Das führt schließlich auf

$$u_1 = \frac{B}{\sqrt{12L}} \tag{44}$$

Die Gleichungen (43) und (44) gelten auch für $L = 1$, d.h. für ein einfaches Merkmal X. In diesem Fall ist $h_1 = 1$. Bei gleichen Gewichten h_l ist wieder $u_2 = Nu_1$. Die charakteristischen Standardunsicherheiten hängen also im Wesentlichen nur von der Bewertungsskala und von der Anzahl L der Merkmale ab, aus denen sich das Merkmal Z zusammensetzt. Bei $L = 1$ ist u_1 der durch die Skalenstufung allein bewirkte Unsicherheitsbeitrag.

Damit der Informationsmangel vertretbar bleibt, muss $u(z)$ wesentlich kleiner als u_2 sein. Diese Forderung lässt sich in die Informationsbedingung

$$u(z) < f \cdot u_2 \tag{45}$$

kleiden mit einem geeignet zu wählenden Faktor f, der kleiner als 1 sein muss, aber auch größer als $1/N$, weil mit $f \leq 1/N$ die Bedingung $u(z) < u_1$ entstehen würde. Diese ist nicht erfüllbar, weil u_1 schon die kleinste mögliche Standardunsicherheit ist. Welcher Wert für f gewählt werden sollte, wird im Rahmen der Evaluierung untersucht (Abschnitt 6.1.2). Danach ist ein Wert von etwa 0,5 bis etwa 0,75 sinnvoll. Wünschenswert wäre ein Wert für f, der generell für alle AC empfohlen werden könnte. Leider scheint es einen solchen Wert jedoch nicht zu geben. Es muss hier noch einmal betont werden: eine zu große Unsicherheit bedeutet nicht, dass die Ausprägung des Merkmals beim Teilnehmer zu sehr schwankt, sondern nur, dass der Kenntnismangel über die tatsächlich vorliegende Ausprägung zu groß ist.

Je mehr Merkmale zum Ergebnis eines zusammengesetzten Merkmals beitragen, desto kleiner wird im Allgemeinen die zugehörige Standardunsicherheit. Allerdings gilt hierbei die Faustregel, wie man an Gleichung (44) sieht, dass eine Halbierung der Unsicherheit die vierfache Anzahl L an Merkmalen und einen dementsprechend höheren Aufwand erfordert. Für eine ausreichend kleine Standardunsicherheit des Gesamtmerkmals eines Teilnehmers genügen aber meist etwa $L = 10$ wichtige Merkmale im AC. Dann ist $u_1 = 0{,}091$, wenn $B = 1$, wie meist üblich. Der Unsicherheitsbereich eines Merkmals liegt immer innerhalb der Skalengrenzen und das zugehörige Ergebnis in der Mitte des Unsicherheitsbereichs. Bei großer Unsicherheit tendiert das Ergebnis deshalb zur Mitte der Bewertungsskala. Ähnlich dazu tendiert auch in der konventionellen Statistik der Mittelwert der Bewertungen des Merkmals zur Skalenmitte bei starker Streuung der Bewertungen. Die aufgezeigte Tendenz ist bei kritischer Betrachtung des Ergebnisses zu beachten, insbesondere beim Vergleich zu einem anderen, zwar weniger guten Ergebnis, zu dem aber eine kleinere Unsicherheit gehört. Das letztere Ergebnis sollte wegen dieser kleineren Unsicherheit vorgezogen werden, wenn sich die beiden Ergebnisse nicht signifikant unterscheiden (Abschnitt 3.3.3). Siehe hierzu als Beispiel

in Anhang B Anlage 2 zum Merkmal 2 „Fachkenntnisse“ die sich nicht signifikant unterscheidenden Ergebnisse der vier Teilnehmer nahe der Skalenmitte mit relativ großen zugehörigen Standardunsicherheiten: Das Ergebnis von Teilnehmer 1 ist zwar weniger gut als die anderen, trotzdem sollte es vorgezogen werden, weil zu ihm die kleinste Standardunsicherheit gehört.

4.6 Implementierung des Auswerteverfahrens

Die Durchführung des Auswerteverfahrens nach Abschnitt 4.4 mittels EDV erfordert ein Computer-Programmsystem. In diesem Abschnitt wird aufgezeigt, welche Struktur und wesentlichen Funktionen ein solches Programmsystem aufweisen sollte. Beispiel ist das in Anhang B beschriebene Programm QWAHL. Dieses ist ein reines Experimentier- und Demonstrationsprogramm, das zwar die meisten wichtigen Funktionen umfasst und daher durchaus in einem AC benutzt werden kann, jedoch für den Routineeinsatz in der Praxis noch nicht komfortabel genug ist, insbesondere hinsichtlich der Dateneingabe. Auch sind nicht alle Möglichkeiten, die das Auswerteverfahren bietet, z.B. das Akzeptanzkriterium (Abschnitt 4.5.1) und die Informationsbedingung nach Gleichung (45), vollständig implementiert. Die mit diesem Programm gewonnenen Erfahrungen sind in die folgenden Empfehlungen eingeflossen.

4.6.1 Allgemeine Struktur des Programmsystems

Das Programmsystem muss von vornherein so allgemein konzipiert werden, dass es in jedem AC-analogen Urteils- und Entscheidungsprozess benutzt werden kann. Dies ist die wichtigste Anforderung, die zu erfüllen ist. Sie erfordert eine zunächst ganz abstrakte Struktur des Programmsystems, die für ein aktuelles AC zu konkretisieren ist. Deshalb muss das Programmsystems bei seiner Anwendung zuallererst erfahren, wie das auszuwertende AC aufgebaut ist. Das Programmsystem kann aus selbständigen, aufeinander abgestimmten Programmen bestehen, die wesentliche Teilaufgaben wahrnehmen, z.B. die Dateneingabe, die eigentlichen Berechnungen oder die Präsentation der Ergebnisse, und die bei Bedarf aufgerufen werden, oder aber aus Programmteilen für jene Aufgaben unter der Regie eines Menüprogrammteils. Auch eine gemischte Form des Programmsystems kann sehr zweckmäßig sein wie bei dem Beispielprogramm QWAHL. Hierbei wird für die Erstellung einer Eingabedatei ein gängiges Editierprogramm benutzt. Das Programm QWAHL liest diese Daten, wertet sie aus und präsentiert die Ergebnisse je nach Fragestellung über ein Menü. Unabhängig davon, wie das System programmtechnisch organisiert wird, sollte es funktionell wie folgt gegliedert werden.

4.6.2 Eingabe und Vorbereitung der Daten

In der Vorbereitungsphase des AC wird zunächst ein Programmteil benötigt für die Erstellung und Bearbeitung einer ersten, geeignet zu strukturierenden Eingabedatei mit allen zum Aufbau des aktuellen AC erforderlichen allgemeinen Daten (siehe hierzu das Beispiel einer Eingabedatei in Anhang B, Anlage 1). Im Wesentlichen sind dies die folgenden Daten:

1) Titel des AC,
2) Namen bzw. Titel und Anzahlen der Teilnehmer, Beobachter, Merkmalsgruppen, Merkmale, Verhaltensanker und Übungen,
3) Matrix der Merkmale und Übungen mit allen Gewichten und Zuordnungen der Verhaltensanker zu den Merkmalen und der Merkmale zu den Merkmalsgruppen,
4) Daten zur Festlegung der Bewertungsskala,
5) globale relative Standardunsicherheiten der Gewichte,
6) Akzeptanzgrenzwerte der Merkmale,
7) bereits bekannte Ergebnisse und zugehörige Standardunsicherheiten zu Merkmalen der Teilnehmer, z.B. aus einer Vorauswahl,
8) Unsicherheitskennwerte zu Merkmalen, wenn sie aus vorangegangen Anwendungen des AC vorliegen (Abschnitt 6.2),
9) Angaben für die Identifizierung der Bewertungsformularblätter.

Dieser Programmteil sollte auch den Ausdruck der Bewertungsformularblätter wahrnehmen.

Sehr wichtig ist ein weiterer Programmteil für die Erstellung und Bearbeitung einer zweiten, geeignet zu strukturierenden Eingabedatei mit allen im AC gewonnenen Bewertungsdaten (siehe hierzu ebenfalls das Beispiel einer Eingabedatei in Anhang B, Anlage 1, wobei allerdings die beiden hier genannten Eingabedateien vereinigt sind). Seine wichtigste Funktion ist das schnelle Erfassen der Daten von den im AC ausgefüllten Bewertungsformularblättern mit Ablage eines identifizierbaren Datensatzes zu jedem Blatt in der Datei. Die Dateneingabe kann über Tastatur oder Maus in eine Maske am Bildschirm erfolgen, schneller jedoch durch Einscannen, wofür allerdings ein spezielles Leseprogramm wie TELEform sowie ein Flachbett-, Hand- oder Einzugscanner erforderlich sind. Dieser Programmteil sollte auch die Blätter auf mangelhafte Angaben und Vollständigkeit prüfen und die Korrektur der Bewertungsdaten ermöglichen.

Ein dritter Programmteil sollte für die Verwaltung aller Ein- und Ausgabedateien zu unterschiedlichen AC sowie für die Erstellung und Bearbeitung einer Steuerdatei vor-

gesehen werden, die für die Evaluierung einer AC-Serie benötigt wird, wenn dabei eine ganze Reihe von Ein- und Ausgabedateien auszuwerten sind.

4.6.3 Durchführung der Auswertung

Der zentrale Programmteil für die eigentliche Auswertung nach dem Verfahren von Abschnitt 4.4 sollte die folgenden Funktionen wahrnehmen:

1) Bereitstellen der Daten aus den beiden Eingabedateien,
2) Normieren der Gewichte (Abschnitt 4.3.1),
3) für jeden Teilnehmer Berechnen der Ergebnisse und zugehörigen Standardunsicherheiten und Ränge zu allen Merkmalen, Merkmalsgruppen und weiteren interessierenden zusammengesetzten Merkmalen, z.B. zum Gesamtmerkmal (Abschnitt 4.4),
4) Ablegen der errechneten Daten in eine Ausgabedatei für die Visualisierung (Abschnitt 4.6.4) und Evaluierung.

Der Aufwand für die Berechnungen erweist sich als gering im Vergleich zum Aufwand für die Handhabung und Verwaltung aller Daten. Die Berechnungen können sehr schnell erfolgen, deshalb kann die Ausgabedatei temporär sein, sie kann bei Bedarf schnell neu erstellt werden. Für die Evaluierung einer AC-Serie ist es jedoch sinnvoll, die Ausgabedateien der AC-Serie zu archivieren. (Siehe hierzu das Beispiel einer Ausgabedatei in Anhang B, Anlage 2)

Ein weiterer Programmteil dient der Evaluierung einer AC-Serie nach Anwendung des AC auf viele Teilnehmer (siehe auch Kapitel 5 und 6 und Anhang A). Er sollte die folgenden Funktionen aufweisen:

1) Bereitstellen der Daten aus den Ausgabedateien der AC-Serie,
2) Berechnen und Interpretieren der Unsicherheitskennwerte der Übungen bezüglich der Merkmale, der Merkmale selbst sowie der Beobachter (Abschnitt 6.2),
3) Berechnen und Interpretieren der Korrelationskoeffizienten zur Konstruktvalidität und der teilnehmerbezogenen Korrelationskoeffizienten (Abschnitt 6.3),
4) Aufstellen einer Statistik der Korrelationskoeffizienten (Abschnitt 6.3),
5) Ablegen der errechneten Daten in eine Evaluierungs-Ausgabedatei für die Visualisierung.

4.6.4 Ausgabe und Visualisierung der Ergebnisse

Schließlich wird noch ein Programmteil für die besonders wichtige Visualisierung aller Ergebnisse am Bildschirm für die Beobachterkonferenz oder durch Ausdruck benötigt. Eine anschauliche und überzeugende Visualisierung kann für die Akzeptanz des Verfahrens entscheidend sein. Die Visualisierung sollte sowohl in graphischer als auch in tabellarischer Form nach unterschiedlichen praxisrelevanten Fragestellungen menügeführt erfolgen, d.h. über ein Menü sollte ein schnelles Umschalten auf die Darstellung zu einer anderen Fragestellung möglich sein (Beispiel siehe Anhang B). Der Programmteil hat die folgenden Aufgaben:

1) Bereitstellen der Daten aus den Ausgabedateien,
2) Darstellen aller Ergebnisse zu jedem Teilnehmer (Teilnehmerprofil),
3) Vergleichen der Teilnehmer hinsichtlich ihrer Ergebnisse und ihrer Ränge zum Gesamtmerkmal, zu jeder einzelnen Merkmalsgruppe, zu jedem einzelnen Merkmal und zu jeder einzelnen Übung,
4) Generieren von Warnmeldungen bei zu großer Unsicherheit, Verletzung des Akzeptanzkriteriums oder der Informationsbedingung bei Merkmalen sowie bei anderen Auffälligkeiten,
5) Darstellen der Ergebnisse der Evaluierung,
6) Generieren von Meldungen bei Auffälligkeiten hinsichtlich der Interpretation der Unsicherheitskennwerte und Korrelationskoeffizienten (Abschnitte 6.2 und 6.3).

5 AC-Serien zur Evaluierung des Auswerteverfahrens

5.1 Allgemeines zur Evaluierung

Zweck der Kapitel 5 und 6 ist es, die Anwendbarkeit des in Kapitel 4 entwickelten Auswerteverfahrens durch den praktischen Einsatz bei drei AC-Serien zu prüfen. Es soll gezeigt werden, was die Berücksichtigung der Unsicherheit einerseits für die Bewertung der Teilnehmer und andererseits für die Evaluierung der Übungen, Merkmale und Beobachter in der Praxis bietet. In diesem Kapitel 5 werden die AC-Serien beschrieben. Durch Vergleich der Teilnehmer-Ergebnisse des Auswerteverfahrens mit denen der Beobachterkonferenzen wird das Verfahren danach in Kapitel 6 evaluiert. Außerdem werden Unsicherheitskennwerte der Übungen, Merkmale und Beobachter sowie Korrelationskoeffizienten zur Konstruktvalidität der Merkmale in den Übungen untersucht, die Korrelationskoeffizienten sowohl nach der konventionellen als auch nach der Bayes'schen Statistik. Insbesondere wird geprüft, ob die für die Berechnung der Unsicherheiten wichtige Voraussetzung erfüllt ist, dass die Merkmale nicht teilnehmerbezogen korreliert sind (Abschnitt 3.5.2).

Betont wird, dass es in erster Linie darum geht, das in Kapitel 4 entwickelte Auswerteverfahren für ein AC mit Berücksichtigung der Unsicherheit zu evaluieren, n i c h t jedoch die AC-Serien hinsichtlich ihrer Konstruktvalidität (Kleinmann, 1997), d.h. hinsichtlich der Frage, ob die in einem AC vorgesehenen Merkmale und Übungen sinnvoll sind. Trotzdem wird gezeigt, dass die in Abschnitt 6.2 betrachteten neuen Unsicherheitskennwerte der Übungen und Merkmale auch zur Klärung dieser Frage beitragen können.

5.2 Übersicht zu den AC-Serien für die Evaluierung

Zu den drei AC-Serien, die für die Evaluierung des Auswerteverfahrens mit Berücksichtigung der Unsicherheit herangezogen werden, siehe die folgende Tabelle 5.1 mit den Übersichtsdaten sowie die Tabellen 5.2 und 5.3, in denen die Matrizen der Merkmale und Übungen der AC-Serien aufgeführt sind. In diesem Kapitel 5 sind die Beobachter nur als bewertende einzelne Personen zu verstehen. Nach Abschnitt 2.2.3 sind zwar in allgemeinerer Sicht bewertende Teams sowie Tests ebenfalls als „Beobachter" aufzufassen, jedoch bleibt es hier bei den Benennungen Team und Test.

Bei der ersten AC-Serie (AC-Serie JA) handelt es sich um ein eintägiges Auswahl-AC zur Einstellung in den Justizvollzugsdienst. Das AC wurde 14-mal mit insgesamt 78 Teilnehmern und insgesamt 4 Beobachtern durchgeführt. Diese AC-Serie JA ist die am klarsten konzipierte und am konsequentesten durchgeführte unter den drei betrachteten AC-Serien. Die zweite AC-Serie (AC-Serie JB) ist vom gleichen Typ wie die erste und diente der gleichen Zielsetzung. Sie unterscheidet sich in ihrer Konstruktion, d.h. in ihren Merkmalen und Übungen, nur geringfügig von der AC-Serie JA (Tabelle 5.2), in der Durchführung allerdings im Wesentlichen in der größeren Anzahl von Beobachtern. Das AC fand 10-mal mit insgesamt 54 Teilnehmern und 12 Beobachtern statt.

Tabelle 5.1: Übersichtdaten der AC-Serien zur Evaluierung

(Beobachter sind hier bewertende Personen)

AC-Serie:	**JA**	**JB**	**ST**
Typ	Auswahl-AC	Auswahl-AC	Potenzial-AC
Jahr der Durchführung	1999	2000	2000
Anzahl der AC	**14**	**10**	**3**
Anzahl der Merkmale	10	10	8
Anzahl der Übungen (davon Tests)	6 (3)	6 (3)	7 (2)
Anzahl der Merkmale pro Übung,			
bewertet durch Beobachter	4	3 oder 4	3
bewertet durch Tests	1	1	1
Anzahl der Verhaltensanker pro Merkmal,			
bewertet durch Beobachter	0	0	4 bis 6
bewertet durch Tests	0	0	0 oder 4
Gesamtanzahl der Teilnehmer	**78**	**54**	**25**
Anzahl der Teilnehmer pro AC	2 bis 8	3 bis 10	7, 7, 11
Gesamtanzahl der Beobachter	4	12	13
Anzahl der Beobachter pro AC	4	4 bis 6	7 bis 9
Gesamtanzahl der Beobachterteams	2	unbekannt	7
Anzahl der Beobachterteams pro AC	2	meist 2, z.T. unbekannt	2 oder 3
Anzahl der Beobachter pro Team	2	1 bis 6, meist 2 oder 3	3 oder 4
Bewertungsskala (Stufenhöhe; bester Wert)	1 bis 6 (1; 1)	1 bis 6 (1; 1)	1 bis 5 (0,5; 5)
Einzelbewertung durch Tests und ...	Beobachter	Beobachter	Teams
als Minimum und Maximum	ja	ja	nein

Für die Evaluierung einer AC-Serie hinsichtlich der Konstruktvalidität und der teilnehmerbezogenen Korrelation wie in Abschnitt 6.3 ist es nötig, dass sie auf möglichst viele Teilnehmer angewendet wird, damit interessierende systematische Effekte sich gegenüber zufälligen genügend stark hervorheben. Darum wurden für diesen Zweck die sehr ähnlichen AC-Serien JA und JB zu einer weiteren AC-Serie JC mit 132 Teilnehmern vereinigt, die allerdings nicht als eigenständig und unabhängig von jenen angesehen werden kann (Abschnitt 5.3.4).

Bei der dritten AC-Serie (AC-Serie ST) aus der Praxis handelt es sich um ein zweitägiges Potenzial-AC in einem Unternehmen. Das AC wurde 3-mal mit insgesamt 25 Teilnehmern und 13 Beobachtern durchgeführt. Zu den Merkmalen und Übungen dieser AC-Serie ST siehe Tabelle 5.3.

Die betrachteten AC-Serien JA, JB, JC und ST werden im Folgenden oft nur mit dem entsprechenden Kürzel JA, JB, JC bzw. ST bezeichnet.

5.3 AC-Serien JA und JB zur Einstellung in den Justizvollzugsdienst

5.3.1 Teilnehmer und Beobachter

Die AC-Serie JA wurde im Jahr 1999 von Mitarbeitern der Justizvollzugsschule des Landes Niedersachsen in Wolfenbüttel konzipiert und durchgeführt. Alle Teilnehmer von JA waren Bewerber um die Ausbildung für den gehobenen Justiz- und Verwaltungsdienst des Landes Niedersachsen, bei der ein Fachhochschulstudium absolviert wird. Von den insgesamt 78 Teilnehmern waren 40 männlich und 38 weiblich. Die meisten von ihnen waren zwischen 20 und 28 Jahren alt und hatten das Abitur abgelegt. Einige dieser Teilnehmer hatten eine Ausbildung abgeschlossen und schon Berufserfahrung gesammelt. 8 Teilnehmer (2 Frauen und 6 Männer) waren im Alter zwischen Mitte und Ende 30 und hatten bereits eine Ausbildung für den mittleren Justiz- und Verwaltungsdienst durchlaufen und dort Berufserfahrung gesammelt. Sie strebten die weiterführende Ausbildung für den gehobenen Dienst mit dem Ziel eines beruflichen Aufstiegs an.

Bei den 4 Beobachtern von JA handelte es sich um 2 Frauen und 2 Männer im Alter zwischen Mitte 30 und Ende 50. Je 2 Beobachter bildeten ein Beobachterteam, das in allen 14 einzelnen AC von JA zusammenblieb. Zwei der Beobachter waren Diplom-Psychologen, die je ein Beobachterteam moderierten, bei den beiden anderen Beobachtern handelte es sich um den Leiter und den Geschäftsleiter einer Justizvollzugsanstalt.

Die im Jahr 2000 durchgeführte AC-Serie JB diente demselben Zweck wie JA. Allerdings sind zu den insgesamt 54 Teilnehmern von JB wegen der starken Anonymisierung der vom Veranstalter für die Auswertung überlassenen Daten keine Angaben zur Alters- und Geschlechtsverteilung und zur Vorbildung verfügbar. Es kann jedoch von einer JA ähnlichen Struktur der Teilnehmerschaft ausgegangen werden.

Bei den 12 Beobachtern von JB handelte es sich um 7 Frauen und 5 Männer. In den 10 einzelnen AC von JB mit jeweils 3 bis 10 Teilnehmern wurden jeweils 4 bis 6 Beobachter eingesetzt. Meist bildeten je 2 oder 3 Beobachter ein Beobachterteam, jedoch gab es auch Teams mit 1, 4, 5 und 6 Beobachtern. Die Zusammensetzung der Teams wechselte, in einigen der einzelnen AC auch von Teilnehmer zu Teilnehmer ohne erkennbare Systematik. Über Altersverteilung, Ausbildung und Stellung der Beobachter liegen keine Angaben vor.

5.3.2 Merkmale und Übungen, Bewertungsskala

Bei JA und JB wurden in 6 Übungen insgesamt 10 Merkmale bewertet. Die Matrix in Tabelle 5.2 zeigt, welche Merkmale mit welchen Übungen erhoben wurden. Die 3 ersten Übungen in dieser Tabelle sind Tests. Sie wurden im Sinne von Abschnitt 2.2.3 ebenfalls als „Beobachter“ aufgefasst. Ihre Ergebnisse wurden direkt als Bewertungen genommen. Die Beobachterpersonen wurden nur bei den anderen 3 Übungen eingesetzt. Die Merkmale dieser Übungen wurden operationalisiert, jedoch wurden die Komponenten dieser Operationalisierung, die Verhaltensanker, nicht selbst bewertet. Bei jeder dieser Übungen wurden 3 oder 4 Merkmale gleichzeitig beobachtet. Bei der Konstruktion des AC wurden im Vorfeld generell keine Gewichte für die einzelnen Merkmale festgelegt. Deshalb wurden alle Gewichte der Merkmale gleich 1 gesetzt, sofern diese in den Übungen zu bewerten waren, und sonst gleich 0.

Die Bewertungen des Merkmals 4 „Einfühlungsvermögen“ in der Übung 5 „Präsentation“ von JA erweckten Zweifel hinsichtlich der Konstruktvalidität der Übung 5 zum Merkmal 4. Deshalb wurde in JB hier nicht mehr bewertet. Siehe hierzu auch Abschnitt 6.2.4.

Die Bewertung erfolgte auf einer ganzzahlig gestuften Bewertungsskala von 1 bis 6 mit der Stufenhöhe 1, wobei den Schulnoten entsprechend die 1 die beste mögliche Bewertung war.

Tabelle 5.2: Matrix der Merkmale und Übungen eines AC im Justizvollzugsdienst (AC-Serien JA, JB und JC)

•: Merkmal in der Übung bewertet (Gewicht gleich 1, sonst 0)

Übungen (U): / Merkmale (M) M-Nr. U-Nr.:	Raven APM 1	Konzentrationstest d2 2	Wegeaufgabe 3	Konfliktrollenspiel 4	Präsentation 5	Selbstpräsentation, Interview 6
1 Allgemeine intellektuelle Leistungsfähigkeit	•					
2 Konzentrationsfähigkeit		•				
3 Organisationsfähigkeit			•			
4 Einfühlungsvermögen				•	• nur JA	•
5 Konfliktfähigkeit				•		
6 Pädagogisches Geschick, Durchsetzungsfähigkeit				•		
7 Formale Kommunikationsfähigkeit				•	•	
8 Reflexionsvermögen					•	•
9 Präsentationstechnik					•	•
10 Initiative, Motivation						•

5.3.3 Ablauf

Alle Beobachter wurden vor der Durchführung von JA und JB in einer eintägigen Schulung in die Grundlagen der Beobachtung, Merkmale und angewendeten Übungen eingeführt. Sie wurden auch darauf hingewiesen, dass sie sich beim Bewerten, anders als sonst üblich, jeweils nicht auf eine Bewertung festzulegen brauchen, sondern ihre subjektive Unsicherheit beim Bewerten aufgrund ihrer Beobachtungen auch durch die Angabe eines Bewertungsbereichs in Form einer nach ihrer Ansicht noch vernünftigen maximalen und minimalen Bewertung zum Ausdruck bringen können. Wenn aufgrund ihrer Beobachtungen des Verhaltens eines Teilnehmers bei einer Übung nach ihrem subjektiven Eindruck sowohl eine Bewertung 2 oder 3 als auch noch eine 4 realistisch infrage käme, sollten sie einen Bereich von 2 bis 4 angeben. Wenn sie sich jedoch beim Bewerten sehr sicher wären, sollten sie nur eine Bewertung angeben.

Die Beobachter der Teams nahmen im Anschluss an jede Übung g e t r e n n t voneinander jeweils die Bewertung der Merkmale vor. Auf den Bewertungsformularen waren zu jedem Merkmal die Verhaltensanker aufgeführt, die jedoch selbst nicht zu bewerten waren.

Vor Beginn jedes einzelnen AC von JA und JB wurde den jeweiligen Teilnehmern beschrieben, auf was allgemein geachtet wird. Die Merkmale und die dazugehörigen Verhaltensanker wurden allerdings nicht im Einzelnen vorgestellt. Die Teilnehmer wurden den Beobachterteams bei der Durchführung der Übungen so zugeteilt, dass jeder Teilnehmer von jedem Beobachter mindestens einmal bewertet wurde.

Nach Abschluss aller Übungen wurden in einer Beobachterkonferenz alle Bewertungen zusammengetragen, eine Rangfolge der Teilnehmer gebildet (nicht im AC 9 von JB) und eine Entscheidung bezüglich der Einstellung in den gehobenen Justiz- und Verwaltungsdienst getroffen. Anschließend wurde mit jedem Teilnehmer ein Rückmeldegespräch geführt, bei dem ihm die Entscheidung mitgeteilt und über die beobachteten Stärken und Schwächen gesprochen wurde.

JA und JB wurden also so durchgeführt, wie es auch sonst in der Unternehmenspraxis üblich ist. Neu war nur die Methode der Bewertung, bei der die Beobachter aufgefordert waren, auch ihre Unsicherheit beim Bewerten durch Angabe eines Bewertungsbereichs zum Ausdruck zu bringen.

5.3.4 Vereinigung der AC-Serien JA und JB zur AC-Serie JC

Für die Evaluierung einer AC-Serie hinsichtlich der Konstruktvalidität sind Unsicherheitskennwerte und Korrelationskoeffizienten zu berechnen (Abschnitte 6.2 und 6.3). Diese hängen aber nicht allein systematisch von der zu evaluierenden Konstruktion der Merkmale und Übungen ab, sondern auch von den Zufälligkeiten der Bewertungen der Teilnehmer durch die Beobachter. Dies gilt in starkem Maße vor allem für die Korrelationskoeffizienten. Die Zufälligkeiten treten jedoch statistisch in den Hintergrund, wenn die AC-Serie auf genügend viele Teilnehmer angewendet wird.

Aber was heißt „genügend viele Teilnehmer“? Genügen die 78 oder 54 Teilnehmer von JA und JB zum Unterdrücken der zufälligen Einflüsse? Um dies zu klären, könnte man die zu evaluierende AC-Serie in zwei Teilserien mit je der halben Teilnehmerzahl aufteilen, die Teilserien einzeln evaluieren und schließlich sich entsprechende Ergebnisse miteinander vergleichen (Split-half-Evaluation). Das ist sinnvoll, hat aber den Nachteil, dass gerade in den Teilserien mit viel kleinerer Teilnehmerzahl die Zufälligkeiten noch stärkeren Einfluss haben als in der ursprünglichen AC-Serie.

Es gibt noch eine sich hier anbietende bessere Möglichkeit. JA und JB sind in ihrer Konstruktion nahezu gleich, in JA wurde lediglich Merkmal 4 „Einfühlungsvermögen“ in Übung 5 „Präsentation“ gegenüber JB zusätzlich bewertet (Tabelle 5.2). Deshalb

können die Teilnehmer von JA, wenn diese zusätzlichen Bewertungen weggelassen werden, auch als Teilnehmer von JB aufgefasst werden. Auf diese Weise wurde eine weitere AC-Serie JC mit 132 Teilnehmern als Vereinigung von JA und JB gebildet. JB ist dadurch Teilserie von JC, JA mit Einschränkung ebenfalls. JC ist keine eigenständige AC-Serie, sie hat n u r Sinn für die Evaluierung von JA und JB. Die Ergebnisse des Auswerteverfahrens zu den Teilnehmern von JB ändern sich in JC nicht, die zu den Teilnehmern von JA ändern sich zwar aufgrund der weggelassenen Bewertungen, jedoch ist dies nicht relevant. Deshalb wurden sie in JC nicht neu berechnet.

5.4 AC-Serie ST zur Potenzialbeurteilung im Unternehmen

5.4.1 Teilnehmer und Beobachter

Die AC-Serie ST wurde im Jahr 2000 von einem internationalen Logistik-Konzern durchgeführt und von einer externen Unternehmensberatung konzipiert und moderiert. Alle Teilnehmer von ST waren Mitarbeiter des Konzerns und bewarben sich für ein vom Konzern finanziertes und exklusiv entwickeltes Studium der Betriebswirtschaft an einer Wirtschaftsakademie. Das AC war darauf ausgerichtet, festzustellen, ob ein Teilnehmer für dieses Studium empfohlen werden kann oder ob für ihn andere Weiterbildungsmaßnahmen geeigneter erscheinen.

Von den insgesamt 25 Teilnehmern von ST waren 19 männlich und 6 weiblich. Sie waren im Alter zwischen Mitte 20 und Mitte 30, verfügten über eine abgeschlossene Berufsausbildung und hatten einige Jahre Berufserfahrung. Die meisten Teilnehmer hatten das Abitur abgelegt.

Bei den Beobachtern von ST handelte es sich um Mitarbeiter der Personalbereiche sowie um operativ tätige Führungskräfte aus den verschiedenen Geschäftsbereichen des Konzerns. Unter den insgesamt 13 Beobachtern waren 4 Frauen und 9 Männer im Alter zwischen Ende 20 und Mitte 50. An jedem einzelnen der 3 AC von ST nahmen 7 bis 9 Beobachter teil, einige Beobachter nur an einem dieser AC. 3 bis 4 Beobachter bildeten in jedem dieser einzelnen AC jeweils ein Team. Sie legten nach den Übungen g e m e i n s a m ihre Bewertungen nieder. Im Sinne vom Abschnitt 2.2.3 sind also die Teams als die „Beobachter“ aufzufassen, ebenso die 2 Tests, nicht jedoch die einzelnen Beobachterpersonen der Teams (Abschnitte 5.2 und 5.4.2). Insgesamt gab es 7 Teams, 2 oder 3 davon in jedem der 3 einzelnen AC. Jedes Team wurde von einer Diplom-Psychologin oder einem Diplom-Psychologen moderiert.

Tabelle 5.3: Matrix der Merkmale und Übungen eines AC im Unternehmen (AC-Serie ST)

•: Merkmal in der Übung bewertet (Gewicht gleich 1, sonst 0).
Dahinter in Klammern: Anzahl der einzeln bewerteten Verhaltensanker, alle mit Gewicht gleich 1

Übungen (U): / Merkmale (M) M-Nr. U-Nr.:	Postkorb 1	Gruppen-diskus-sion 2	Interview 3	Lerntest 4	Kogniti-ver Test 5	Rollen-spiel 1 6	Rollen-spiel 2 7
1 Kognitive Schnelligkeit, Analysevermögen	• (3)				• (4)		
2 Unternehmerisches Denken	• (4)						
3 Lernmotivation, Belastbarkeit			• (6)	• (0)			
4 Leistungsmotivation			• (5)				
5 Langfristmotivation			• (4)				
6 Teamfähigkeit		• (5)				• (5)	• (5)
7 Konfliktfähigkeit, Durchsetzungskraft		• (4)				• (5)	• (5)
8 Überzeugungskraft		• (4)				• (5)	• (5)

5.4.2 Merkmale und Übungen, Bewertungsskala

In ST wurden in 7 Übungen insgesamt 8 Merkmale bewertet. Die Matrix in Tabelle 5.3 zeigt, welche Merkmale mit welchen Übungen erhoben wurden. Die Übungen 4 und 5 in dieser Tabelle sind Tests. Ihre Ergebnisse wurden direkt als Bewertungen genommen, d.h. im Sinne vom Abschnitt 2.2.3 wurden die Tests ebenfalls als „Beobachter“ aufgefasst. Die Beobachterpersonen wurden nur bei den anderen 4 Übungen eingesetzt (Abschnitt 5.4.1). Bei jeder dieser Übungen wurden 2 oder 3 Merkmale gleichzeitig beobachtet. Diese Merkmale wurden durch je 3 bis 6 Verhaltensanker operationalisiert, die jeweils einzeln bewertet wurden. Nach Abschnitt 2.2.5 wurden die Verhaltensanker bei der Auswertung nicht als eigene Merkmale aufgefasst, sondern ihre Bewertungen dem zugehörigen Merkmal zugeordnet. Bei der Konstruktion des AC wurden im Vorfeld generell keine Gewichte für die einzelnen Merkmale festgelegt. Deshalb wurden alle Gewichte der Merkmale und Verhaltensanker gleich 1 gesetzt, sofern diese in den Übungen zu bewerten waren, und sonst gleich 0.

Die Bewertung wurde auf einer gestuften Bewertungsskala von 1 bis 5 vorgenommen, wobei Zwischenstufen der Höhe 0,5 erlaubt waren und die 5 die beste mögliche Bewertung war. Es handelte sich also tatsächlich um eine 9-stufige Skala mit der Stufenhöhe 0,5.

5.4.3 Ablauf

Vor der Durchführung von ST wurden alle Beobachter in einer eintägigen Schulung in die Grundlagen der Beobachtung, Merkmale und angewendeten Übungen eingeführt. Jeweils 3 bis 4 Beobachter bildeten Teams und nahmen, wie es in der Praxis manchmal gehandhabt wird, nach jeder Übung zunächst jeder für sich und anschließend g e m e i n s a m eine Bewertung der Merkmale vor.

Zur Bewertung der Merkmale wurde auf einem Bewertungsformular mit Verhaltensankern gearbeitet. Jeder Verhaltensanker der operationalisierten Merkmale wurde von den Teams einzeln bewertet. Im Gegensatz zu JA und JB wurden die Teams von ST n i c h t dazu angehalten, jeweils eine minimale und maximale Bewertung abzugeben, um ihre Unsicherheit beim Bewerten auszudrücken. Deshalb liegt für die Verhaltensanker meist nur je eine Bewertung vor. Die Unsicherheit eines Merkmals lässt sich aber über die Verteilung der Bewertungen der zu diesem Merkmal gehörenden Verhaltensanker ermitteln.

Abschließend wurde nach jeder Übung aus den Bewertungen der Verhaltensanker jedes Merkmals eine Gesamtbewertung für das Merkmal gebildet.

Vor Beginn jedes einzelnen AC von ST wurden den jeweiligen Teilnehmern die zu bewertenden Merkmale bekannt gegeben. Die Verhaltensanker wurden jedoch nicht im Einzelnen erläutert. Im Anschluss an das AC wurden mit den Teilnehmern in individuellen Rückmeldegesprächen ihre Ergebnisse besprochen und eine Empfehlung bezüglich der Teilnahme an der Weiterbildung ausgesprochen. Die Teilnehmer erhielten außerdem etwa zwei Wochen nach dem AC ein schriftliches Kurzgutachten über ihre Ergebnisse.

ST wurde also ebenfalls so durchgeführt, wie es auch sonst in der Unternehmenspraxis üblich ist.

6 Evaluierung des Auswerteverfahrens

6.1 Vergleich der Teilnehmer-Ergebnisse des Auswerteverfahrens und der Beobachterkonferenzen

Das in Kapitel 4 eingeführte Auswerteverfahren für ein AC wird hauptsächlich dadurch evaluiert, dass die mit diesem Verfahren gewonnenen Teilnehmer-Ergebnisse der in Kapitel 5 beschriebenen AC-Serien JA, JB und ST mit den entsprechenden Teilnehmer-Ergebnissen aus den Beobachterkonferenzen dieser AC-Serien verglichen werden. Ein Teilnehmer-Ergebnis ist in diesem Kapitel 6 als Ergebnis zum Gesamtmerkmal eines Teilnehmers zu verstehen. Außerdem sollen in diesem Abschnitt die Vorteile deutlich werden, die die Ermittlung der Standardunsicherheiten zu den Teilnehmer-Ergebnissen für die Beobachterkonferenz eines AC bieten kann. Diese Unsicherheiten ermöglichen es, die Qualität der Teilnehmer-Ergebnisse zu begutachten, sie spielen aber auch eine wichtige Rolle bei dem genannten Vergleich.

6.1.1 Berechnung und Angabe der Teilnehmer-Ergebnisse

Die Teilnehmer-Ergebnisse von JA, JB und ST sind in den Tabellen A.1, A.2 bzw. A.3 in Anhang A angegeben. In JC sind JA und JB vereinigt, zu den Teilnehmer-Ergebnissen von JC siehe deshalb die Tabellen A.1 und A.2 und Abschnitt 5.3.4. Zu jedem Teilnehmer gehört eine Tabellenzeile. Die Tabellenzeilen zu den Teilnehmern an einem einzelnen AC der AC-Serien bilden jeweils eine Zeilengruppe, stehen direkt untereinander und sind durch horizontale dünne Linien von den Zeilen anderer Teilnehmer getrennt. Die Tabelleneinträge in einer Zeile haben die folgenden Bedeutungen:

In Spalte 1 steht die laufende Nummer des Teilnehmers innerhalb der AC-Serie, in Spalte 2 die Nummer des einzelnen AC der AC-Serie, in dem der Teilnehmer bewertet wurde, sowie die Nummer des Teilnehmers in diesem einzelnen AC.

In Spalte 3 ist das Teilnehmer-Ergebnis z_K aus der Beobachterkonferenz angegeben (das Kürzel K steht für Konferenz). In den Beobachterkonferenzen wurden Gesamtbewertungen nur zu den einzelnen Merkmalen festgelegt. Das Teilnehmer-Ergebnis z_K wurde als Mittelwert dieser Bewertungen mit gleichen Gewichten berechnet.

Das Ergebnis z_T zum Gesamtmerkmal Z des Teilnehmers (T) nach dem Auswerteverfahren dieser Arbeit steht in Spalte 4, gefolgt in Spalte 5 von der Standardunsicherheit

$u(z_T)$ zu z_T. Die maximale Standardunsicherheit unter den in Spalte 5 aufgeführten Werten ist fett gedruckt. Das Gesamtmerkmal Z wurde aus den bewerteten Merkmalen X_l mit gleichen Gewichten nach Gleichung (37) zusammengesetzt. Die Auswertung erfolgte nach Abschnitt 4.4 und nach Abschluss aller AC-Serien. Die globalen relativen Standardunsicherheiten u_r und u'_r wurden nicht berücksichtigt, also gleich null gesetzt.

In Spalte 6 der Tabellen A.1 und A.2 ist der in der jeweiligen Beobachterkonferenz von JA bzw. JB festgelegte Rang R'_K des Teilnehmers innerhalb des einzelnen AC, in dem der Teilnehmer bewertet wurde, eingetragen. Im AC 9 von JB wurde dieser Rang nicht vergeben. Bei ST ist dieser Rang in Tabelle A.3 nicht aufgeführt, weil er identisch ist mit dem in Spalte 7 stehenden Rang. Die Angabe in Spalte 7 ist der nach den Teilnehmer-Ergebnissen z_K in Spalte 3 gebildete Rang R_K des Teilnehmers. Er wird in Spalte 8 gefolgt vom Rang R_T, der nach den Teilnehmer-Ergebnissen z_T in Spalte 4 entsprechend ermittelt wurde. In Spalte 7 wurde bei gleichen Ergebnissen von n Teilnehmern jedem von ihnen dieselben n aufeinander folgenden Ränge zuwiesen, die Teilnehmer teilen sich also diese Ränge. In Spalte 8 wurde ebenso verfahren, allerdings nur bei auch gleichen Standardunsicherheiten. Sonst wurde bei zwei gleichen Teilnehmer-Ergebnissen demjenigen mit der kleineren Standardunsicherheit der höhere Rang zugesprochen. Sich unterscheidende Ränge in den Spalten 6 und 7 oder in den Spalten 7 und 8 sind in der Bemerkungsspalte 9 durch a bzw. b markiert, jedoch nur dann, wenn bei mehreren zugewiesenen Rängen auf einem Platz in einer Spalte keiner von diesen Rängen mit dem zu vergleichenden Rang in der anderen Spalte übereinstimmt. Hinsichtlich der Vergabe der Ränge ist zu beachten, dass in JA und JB die 1 die beste mögliche Bewertung darstellt, in ST jedoch die 5.

In Spalte 6 ist bei JA und JB außerdem eine in der Beobachterkonferenz ausgesprochene positive Einstellungsempfehlung durch + angezeigt. Bei ST ist entsprechend eine positive Weiterbildungsempfehlung in Spalte 7 durch + markiert. Auch in Spalte 8 ist bei allen drei AC-Serien ein positives Votum dann vermerkt, wenn der Unsicherheitsbereich des Teilnehmers mit den Grenzen $z_T - u(z_T)$ und $z_T + u(z_T)$ vollständig auf der besseren Seite des Akzeptanzgrenzwertes liegt (Abschnitt 4.5.1). Das bedeutet, dass alle vernünftigen Schätzwerte für das Gesamtmerkmal Z des Teilnehmers besser sind als der Akzeptanzgrenzwert. Für diesen wurde der Wert 3 angesetzt. Durch diese Wahl wurden in JA und JB insgesamt 29 positive Voten vergeben, was den 30 Einstellungsempfehlungen der Beobachterkonferenzen in etwa entspricht. In ST bewirkte sie 13 positive Voten, also für etwa die Hälfte der 25 Teilnehmer. Das erscheint vernünftig, steht jedoch der großzügig bemessenen Anzahl von 21 Weiterbildungsempfehlungen der Beobachterkonferenzen von ST gegenüber.

Die Bemerkung c in Spalte 9 (nur bei Teilnehmer 48 von JA) bedeutet, dass z_K nicht im Unsicherheitsbereich zwischen $z_T - u(z_T)$ und $z_T + u(z_T)$ liegt und deshalb kein vernünftiger Schätzwert des Gesamtmerkmals Z im Sinne des Auswertungsverfahrens ist (Abschnitt 6.1.2).

Hier werden nur die Ergebnisse zu den Gesamtmerkmalen der Teilnehmer von JA, JB und ST untersucht und verglichen. Allgemein ist es natürlich möglich, eine AC-Serie nicht nur hinsichtlich der Gesamtmerkmale der Teilnehmer, sondern auf gleiche Weise auch hinsichtlich der Einzel- und Gruppenmerkmale zu analysieren.

6.1.2 Diskussion der Teilnehmer-Ergebnisse und zugehörigen Unsicherheiten

Zuerst wird bei jedem Teilnehmer das Ergebnis z_K aus der Beobachterkonferenz mit dem Ergebnis z_T des Auswerteverfahrens in den Spalten 3 und 4 der Tabellen A.1, A.2 und A.3 verglichen. Es lassen sich zwar Unterschiede zwischen z_K und z_T feststellen, doch ist deren Differenz mit einer einzigen Ausnahme immer dem Betrag nach kleiner als die Standardunsicherheit $u(z_T)$ zu z_T in Spalte 5. Anders ausgedrückt liegt z_K mit einer Ausnahme immer im Unsicherheitsbereich des Gesamtmerkmals Z des Teilnehmers zwischen den Grenzen $z_T - u(z_T)$ und $z_T + u(z_T)$. Daher kann nach Sichtweise dieser Arbeit das Ergebnis z_K aus der Beobachterkonferenz bei allen Teilnehmern außer einem als vernünftiger Schätzwert von Z angesehen werden. In diesem Sinne werden also die Entscheidungen der Beobachterkonferenzen gestützt.

Die einzige Ausnahme ist Teilnehmer 48 von JA. Beim ihm ist außerdem die Standardunsicherheit mit Abstand am größten innerhalb von JA und JB. Ein Blick in die AC-Unterlagen zeigt, dass der Teilnehmer den Konzentrationstest d2 (Übung 2) falsch bearbeitet hat und nicht zur Übung 6 „Selbstpräsentation, Interview“ erschienen ist (Tabelle 5.2). Er konnte deshalb in diesen beiden Übungen nicht bewertet werden. Dementsprechend waren für die Auswertung bei diesen Übungen nach Abschnitt 4.3.3 die Skalengrenzen als minimale und maximale Bewertungen zu verwenden. Das erklärt die große Standardunsicherheit und auch die große Differenz der Ergebnisse z_K und z_T. Die fehlenden Bewertungen sind in z_T berücksichtigt, in z_K jedoch überhaupt nicht.

Es wird nun die Standardunsicherheit $u(z_T)$ zum Teilnehmer-Ergebnis mit den charakteristischen Standardunsicherheiten u_1 und u_2 nach Abschnitt 4.5.2 verglichen. In JA und JB wurden $L = 10$ Merkmale bewertet unter Verwendung der Bewertungskala mit $N = 6$ Stufen der Höhe $B = 1$, in ST waren $L = 8$, $N = 9$ und $B = 0{,}5$ (Abschnitte 5.3.2 und 5.4.2). Daraus ergeben sich nach den Gleichungen (43) und (44) für JA

und JB $u_1 = 0{,}091$ und $u_2 = 0{,}548$ und für ST $u_1 = 0{,}051$ und $u_2 = 0{,}459$. Bei allen Teilnehmern liegt die Standardunsicherheit $u(z_T)$ zwischen diesen Werten von u_1 und u_2. Das sollte auch so erwartet werden und weist zunächst darauf hin, dass widersprüchliche Bewertungen keine merkliche Rolle spielen. Anderenfalls müsste u_2 überschritten sein. Auffällig ist aber, dass die Standardunsicherheiten in ST systematisch größer sind als in JA und JB. Dies ist eine Folge der Streuung der Bewertungen der Verhaltensanker in ST. Offenbar trägt diese Streuung mehr zur Unsicherheit bei als in JA und JB die Art und Weise, für Merkmale direkt minimale und maximale Bewertungen abzugeben. Das kann daran liegen, dass die Beobachter ihre eigene Unsicherheit zu klein einschätzen, vielleicht weil es bisher üblich war, immer nur jeweils eine Bewertung abzugeben und nicht deren zwei. In der Tat kommen Aussagen, bei denen die minimale und maximale Bewertung sich um mehr als eine Stufe unterscheiden, nur selten vor. Es ist aber auch zu erwarten, dass ein Verhaltensanker für sich allein das zugehörige Merkmal nicht genau genug repräsentiert. Das kann bei mehreren Verhaltensankern zu einem Merkmal eine große, aber durchaus realistische Streuung gut erklären.

Die Informationsbedingung $u(z_T) < fu_2$ nach Gleichung (45) sollte auf jeden Fall die größte Standardunsicherheit in JA bei Teilnehmer 48 als zu groß erkennen lassen. Dazu muss $f < 0{,}657$ gewählt werden. Wenn auch die größte Standardunsicherheit in JB bei Teilnehmer 19 als zu groß erkannt werden soll, aber sonst keine weiteren, ist $0{,}475 < f < 0{,}566$ anzusetzen. (Die vorstehenden Zahlenwerte errechnen sich aus den in Spalte 5 der Tabellen A.1 und A.2 angegebenen Standardunsicherheiten.) Hier wäre also $f = 1/2 = 0{,}5$ sinnvoll. In ST würde dieser Wert aber dazu führen, dass nur bei drei von den 25 Teilnehmern die Standardabweichung als nicht zu groß eingestuft wird. Das ist nicht plausibel. Wenn in ST nur die Standardunsicherheiten der Teilnehmer 1 und 18 mit den größten Werten als zu groß erkannt werden sollen, aber sonst keine weiteren, ist $f > 0{,}741$ zu wählen, also z.B. $f = 3/4 = 0{,}75$. Das ist aber für JA und JB ein zu großer Wert. Offenbar ist es nicht möglich, für alle drei untersuchten AC-Serien einen generellen Wert f für zukünftige Anwendungen festzulegen, was ebenfalls an den systematisch unterschiedlich großen Unsicherheiten in den AC-Serien liegt, deren mögliche Ursachen im vorangehenden Absatz genannt wurden. Ob wenigstens bei bestimmten AC-Typen für f ein genereller Wert gewählt werden kann, müssen zukünftige Untersuchungen an vielen weiteren AC-Serien zeigen. Wenn irgendeine andere AC-Serie untersucht wird, sollte f für spätere Anwendungen desselben AC so festgelegt werden, dass die Standardunsicherheiten von etwa 5 % bis 10 % der Teilnehmer als zu groß erkannt werden.

Aus der Diskussion wird klar, dass der Standardunsicherheit bei den Vergleichen eine bedeutende Rolle zukommt.

6.1.3 Diskussion der Teilnehmer-Rangfolgen

Es werden nun die Rangfolgen der Teilnehmer verglichen, zunächst die in Spalte 7 mit denen in Spalte 8 der Tabellen A.1, A.2 und A.3. Diese Rangfolgen bei den Rängen R_K und R_T unterscheiden sich nur wenig. In JA und JB kommt nur je eine Rangvertauschung vor, in ST deren zwei (Bemerkung b in Spalte 9 der Tabellen). In jedem dieser vier Fälle überlappen sich die Unsicherheitsbereiche zu den jeweils beteiligten beiden Teilnehmern. Daraus folgt, dass deren Teilnehmer-Ergebnisse sich nicht signifikant unterscheiden und dass es auch vernünftige Schätzwerte der Gesamtmerkmale dieser beiden Teilnehmer im Überlappungsbereich ihrer Unsicherheitsbereiche in umgekehrter Reihenfolge gibt. Die Rangfolgen in den Spalten 7 und 8 sind daher an keiner Stelle signifikant verschieden.

Anders sieht es aus beim Vergleich der in den Spalten 6 und 7 aufgeführten Rangfolgen der Teilnehmer von JA und JB. Diese Rangfolgen bei den Rängen R'_K und R_K unterscheiden sich erheblich voneinander (Bemerkung a in Spalte 9 der Tabellen A.1 und A.2). Offenbar haben bei der Vergabe der Ränge R'_K in den Beobachterkonferenzen neben den Ergebnissen aus den Übungen auch andere Gesichtspunkte eine Rolle gespielt. In der Tat wurde in den Beobachterkonferenzen von JA nach Aussage der Moderatorin auch die physische und psychische Disposition und Robustheit der Teilnehmer für den rauen Umgang im Justizvollzug intensiv diskutiert, was die Rangvergabe stark beeinflusst habe. Es folgt daraus, dass diese für die Aufgabe wichtige Disposition als ein zusätzliches Merkmal in einer geeigneten Übung hätte bewertet werden müssen.

6.2 Evaluierung der Übungen, Merkmale und Beobachter

In diesem Abschnitt werden *Unsicherheitskennwerte* definiert und berechnet, die auf der Messunsicherheit beruhen und der Evaluierung der Übungen, Merkmale und Beobachter bei einer auf viele Teilnehmer angewendeten AC-Serie dienen sollen. Sie erlauben es, die Konstruktvalidität der Übungen und Merkmale in Ergänzung zu üblichen Methoden zu untersuchen und das Bewertungsverhalten der Beobachter zu begutachten. Um dies zu zeigen, werden die Unsicherheitskennwerte zu den AC-Serien JA, JB, JC und ST berechnet und diskutiert.

6.2.1 Allgemeines zu Unsicherheitskennwerten

Bisher wurde alle in einem AC oder in einer AC-Serie gewonnene Information im Wesentlichen als Aussagen der Beobachter zu den Merkmalen der einzelnen Teilnehmer aufgefasst. Aus anderer Sicht können die Aussagen jedoch auch als Information zu ganz anderen Sachverhalten unter anderen interessierenden Fragestellungen herangezogen werden. In einem solchen Fall muss der Sachverhalt genau beschrieben und auf der Bewertungsskala quantifiziert werden und es ist dem Sachverhalt wie den Merkmalen ein Schätzer Y zuzuordnen. Dann sind aus der gesamten vorliegenden Information die für den Sachverhalt relevanten Aussagen herauszuziehen. Aus diesen lassen sich anschließend nach Abschnitt 3.4.1 mit Hilfe des Bernoulli'schen Prinzips die Wahrscheinlichkeitsverteilung des Schätzers Y aufstellen sowie der beste Schätzwert $y = \mathrm{E}\,Y$ zum Sachverhalt und die zugehörige Standardunsicherheit $u(y) = \sqrt{\mathrm{Var}\,(Y)}$ berechnen. Diese können jedoch auch aus der relevanten Information direkt berechnet werden, wenn die Aussagen in Form minimaler und maximaler Bewertungen gegeben sind.

Das Ergebnis y und die Standardunsicherheit $u(y)$, die immer zusammengehören, hängen also von der Information ab, die für die aktuelle Fragestellung vorliegt und relevant ist. Ändert sich die Fragestellung, so kann andere Information in der Menge der insgesamt vorliegenden Information relevant werden, was auch das Ergebnis und die zugehörige Standardunsicherheit ändert. Das gilt ebenso für die Korrelation zweier Schätzer. Sehr wichtig ist deshalb neben der Information eine ganz genaue Fragestellung. Das ist wie bei einer Statue, die je nach Blickpunkt eine andere Ansicht bietet.

Dies ist das Prinzip des Vorgehens in diesem Abschnitt 6.2 bei der Definition und Berechnung der neuen Unsicherheitskennwerte zum Zweck der Evaluierung der Übungen, Merkmale und Beobachter. Die Unsicherheitskennwerte sind nichts anderes als Standardunsicherheiten zu Sachverhalten, die bestimmte sinnvolle Fragestellungen charakterisieren. Sie sind neue Validitätsmaße, die auf der Bayes'schen Statistik und der Messunsicherheit beruhen.

Nach dem vorstehend skizzierten Prinzip werden Unsicherheitskennwerte zu den folgenden drei Fragestellungen definiert:

1) Wie valide ist eine Übung für die Bewertung eines bestimmten Merkmals? Gibt es Übungen, in denen das Merkmal über alle Teilnehmer und Beobachter einer AC-Serie hinweg systematisch unsicherer bewertet wird als in anderen?

2) Wie genau lässt sich ein Merkmal überhaupt im AC bewerten? Gibt es Merkmale, die über alle Übungen, Teilnehmer und Beobachter einer AC-Serie hinweg systematisch unsicherer bewertet werden als andere?
3) Wie lässt sich das Bewertungsverhalten eines Beobachters im Vergleich zu anderen Beobachtern beurteilen? Gibt es Beobachter, die über alle Übungen, Merkmale und Teilnehmer einer AC-Serie hinweg systematisch unsicherer bewerten als andere?

Der zu definierende Unsicherheitskennwert einer Übung zu einem bestimmten Merkmal bei Fragestellung 1 soll demnach ein Maß für die Konstruktvalidität der Übung für das betrachtete Merkmal bilden und soll die üblichen Maße für diesen Zweck ergänzen, z.B. Korrelationskoeffizienten zur konvergenten und diskriminanten Konstruktvalidität (Abschnitte 6.3.1 und 6.3.2). Der Unsicherheitskennwert eines Merkmals bei Fragestellung 2 ist ein Maß für die Konstruktvalidität des gesamten AC für das Merkmal und der Unsicherheitskennwert eines Beobachters bei Fragestellung 3 ein Maß für das Bewertungsverhalten des Beobachters. Die Unsicherheitskennwerte sind nur sinnvoll im Vergleich mit denen anderer Übungen, Merkmale bzw. Beobachter in demselben AC oder in einem anderen AC unter vergleichbaren Bedingungen, z.B. in einem AC mit teilweise denselben Übungen und derselben Bewertungsskala. Außerdem müssen sie mit den charakteristischen Standardunsicherheiten verglichen werden (Abschnitte 4.5.2 und 6.1.2). Sie sollten erst dann berechnet werden, wenn das AC auf viele Teilnehmer angewendet worden ist, um zufallsbedingte Einflüsse so weit wie möglich gegenüber den interessierenden systematischen Effekten zurückzudrängen.

Das Prinzip der Bildung eines Unsicherheitskennwertes ist natürlich auch auf andere Fragestellungen als die hier betrachteten anwendbar, z.B. auf die Frage, ob Frauen und Männer sich in ihrem Bewertungsverhalten unterscheiden. Allerdings reicht das vorliegende Datenmaterial für eine Evaluierung dieser Frage nicht aus. Der Grund dafür liegt in der Anonymisierung der Daten von JA und JB sowie darin, dass in ST die Beobachter teilweise aus Frauen und Männern gemischte Teams sind. Auch die prognostische Validität eines AC lässt sich mit Hilfe von Unsicherheitskennwerten untersuchen (Abschnitt 7.3). Die Evaluierung dazu muss jedoch zukünftigen Untersuchungen überlassen bleiben, weil dafür stark zeitversetzte Bewertungen derselben Merkmale von Teilnehmern erforderlich sind. Aus den AC-Serien liegen aber nur zeitgleiche Bewertungen vor.

6.2.2 Berechnung der Unsicherheitskennwerte

Für die Berechnung der Unsicherheitskennwerte werden die für die jeweilige Fragestellung relevanten Aussagen A unter allen in einer AC-Serie gewonnenen Aussagen herangezogen und wie in Abschnitt 3.4.1 mit gleichem Gewicht $P(A) = 1/M$ zusam-

mengeführt. Dabei ist M die Anzahl der relevanten Aussagen. Welche Aussagen sind aber für die in Abschnitt 6.2.1 genannten drei Fragestellungen relevant?

Um den Unsicherheitskennwert u_U einer Übung k bezüglich des Merkmals i für Fragestellung 1 zu berechnen, werden die Aussagen aller Beobachter zum Merkmal i in der Übung k zu allen Teilnehmern einer AC-Serie benutzt. Nun ist zu beachten, dass die Ausprägung des Merkmals von Teilnehmer zu Teilnehmer wechselt und deshalb kein Charakteristikum für die Übung sein kann. Darum müssen die in den Aussagen implizit vorhandenen Schätzwerte für das Merkmal bei den Teilnehmern auf die im Folgenden beschriebene Weise eliminiert werden, die Aussagen also zuvor *reduziert*, d.h. auf denselben fiktiven „mittleren“ Teilnehmer bezogen und umgerechnet werden.

Zu diesem Zweck werden zunächst unter den eben herangezogenen Aussagen zum Merkmal i in der Übung k nur diejenigen betrachtet, die zu einem bestimmten Teilnehmer j gehören. Mit diesen Aussagen werden dann nach Abschnitt 3.4.1 der Erwartungswert $x_j = \mathrm{E}\, X_j$ und die Varianz $\mathrm{Var}\,(X_j)$ gebildet. Danach werden alle diese Aussagen auf der Bewertungsskala gemeinsam, d.h. ohne ihre gegenseitige Lage zu verändern, so verschoben, dass der Erwartungswert auf der Skalenmitte zu liegen kommt. Statt der Skalenmitte kann auch ein beliebiger anderer Wert gewählt werden, der aber für alle Teilnehmer derselbe sein muss. Das Folgende hängt davon nicht ab. Wie eben beschrieben, wird für jeden Teilnehmer verfahren. Anschließend wird aus allen so verschobenen Aussagen wieder nach Abschnitt 3.4.1 die Varianz $\mathrm{Var}\,(Y)$ des Schätzers Y zu der betrachteten Fragestellung berechnet. Die Wurzel aus dieser Varianz ist dann der Unsicherheitskennwert. Durch die Verschiebung ändern sich die Varianzen $\mathrm{Var}\,(X_j)$ nicht und alle Erwartungswerte $\mathrm{E}\, X_j$ sind gleich. Und weil auch alle Aussagen und Teilnehmer jeweils mit gleichem Gewicht berücksichtigt werden, folgt daraus, dass die Varianz $\mathrm{Var}\,(Y)$ gleich der über alle Teilnehmer gemittelten Varianz $\mathrm{Var}\,(X_j)$ ist. Damit gilt bei m Teilnehmern

$$u_U = \sqrt{\mathrm{Var}\,(Y)} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} \mathrm{Var}\,(X_j)} \tag{46}$$

Die Unsicherheitskennwerte u_U aller Übungen zu allen Merkmalen bilden eine Matrix. Die hier geschilderte Reduktion ähnelt dem Vorgehen in Abschnitt 3.5.2 bei der teilnehmerbezogenen Korrelation, bezieht sich aber nur auf ein Merkmal, nicht wie dort auf zwei.

Die Unsicherheitskennwerte u_U der Übungen bezüglich der Merkmale können auch, wie in Abschnitt 4.3.1 behandelt, bei einer späteren Anwendung des AC für die Festlegung der Gewichte der Übungen bezüglich der Merkmale herangezogen werden.

Zur Ermittung des Unsicherheitskennwertes u_M eines Merkmals i für Fragestellung 2 werden die Aussagen aller Beobachter zu diesem Merkmal und zu allen Teilnehmern in allen Übungen einer AC-Serie verwendet. Auch dabei müssen die Aussagen zuvor, wie oben beschrieben, reduziert werden. Der Unterschied ist nur, dass hier Aussagen aus allen Übungen, in denen das Merkmal i bewertet wurde, relevant sind. Gleichung (46) gilt analog auch für u_M. Der für die Berechnung von Var (X_j) benötigte Erwartungswert E X_j ist hier der nach Abschnitt 4.4 berechnete beste Schätzwert x für das Merkmal i des Teilnehmers j.

Der Unsicherheitskennwert u_M enthält neben den Unsicherheitsbeiträgen zum betrachteten Merkmal i aus jeder einzelnen Übung k, die durch die Unsicherheitskennwerte u_U charakterisiert werden, auch noch den Unsicherheitsbeitrag, der auf der Streuung der Bewertungen des Merkmals über alle Übungen hinweg beruht. Wird das Merkmal allerdings nur in einer einzigen Übung bewertet, so ist $u_M = u_U$. Da der Unsicherheitskennwert u_M die Unsicherheit eines Merkmals über alle Übungen und Teilnehmer hinweg charakterisiert, kann er in einer späteren Anwendung des AC als Ersatz für die Standardunsicherheit des Merkmals bei einem Teilnehmer dienen, wenn Information zur Ermittlung dieser Standardunsicherheit fehlt oder nicht gewonnen werden kann (Abschnitt 4.3.3).

Der Unsicherheitskennwert u_B eines Beobachters i für Fragestellung 3 erfordert zu seiner Ermittlung alle Aussagen dieses Beobachters zu allen Merkmalen und zu allen Teilnehmern in allen Übungen einer AC-Serie. Wieder müssen die Aussagen zuvor reduziert werden und Gleichung (46) gilt analog auch für u_B. Es gibt hier jedoch zwei sinnvolle Möglichkeiten, das Bewertungsverhalten des Beobachters zu charakterisieren, und daher zwei Arten der Reduktion. Die eine Möglichkeit besteht darin, den Beobachter bei der Berechnung des Unsicherheitskennwertes u_B allein für sich zu betrachten. Bei der anderen Möglichkeit wird der Beobachter zusätzlich hinsichtlich der Abweichungen seiner Bewertungen von den Bewertungen der anderen Beobachter charakterisiert. Der aus dieser zweiten Möglichkeit folgende Unsicherheitskennwert wird mit u'_B bezeichnet. Dieser ist natürlich nicht kleiner als u_B, weil er auch noch den Unsicherheitsbeitrag aufgrund der Abweichungen enthält.

Bei der ersten Art der Reduktion wird jede einzelne Aussage des Beobachters, zu welchem Merkmal und Teilnehmer auch immer, als Aussage zu einem anderen Teilnehmer aufgefasst. Danach wird so verfahren, wie oben ausgeführt. Es sind also ebenso viele „Teilnehmer“ wie Aussagen vorhanden und jede einzelne Aussage wird so verschoben, dass die Mitte zwischen der minimalen und der maximalen Bewertung der Aussage auf der Skalenmitte zu liegen kommt. Gleichung (46) gilt nun analog auch für u_B, aber m

ist die Anzahl der Aussagen und Var (X_j) die Varianz einer Aussage j. Nach Abschnitt 3.4.1 ergibt sich Var $(X_j) = (t_k - t_{j'-1})^2/12$. (Das dortige j ist hier mit j' bezeichnet, um es von dem j an dieser Stelle zu unterscheiden.) Der Unsicherheitskennwert u_B charakterisiert somit die Differenz zwischen minimaler und maximaler Bewertung in allen Aussagen des Beobachters.

Bei der zweiten Art der Reduktion wird jede einzelne Aussage des Beobachters zu einem Merkmal und Teilnehmer in einer Übung in Beziehung gesetzt zu dem sich nach dem Auswerteverfahren von Abschnitt 4.4 ergebenden und schon bei u_U verwendeten besten Schätzwert x_j dieses Merkmals bei diesem Teilnehmer in der jeweiligen Übung. Die Aussage wird gemeinsam mit x_j verschoben bis x_j auf der Skalenmitte zu liegen kommt. Dann wird wieder die Varianz berechnet und daraus die Wurzel gezogen, was den Unsicherheitskennwert u'_B ergibt. Wieder gilt Gleichung (46) analog und m ist die Anzahl der Aussagen in diesem Fall. Allerdings ist jetzt im Gegensatz zu Var (X_j) im vorangehenden Absatz

$$\text{Var}\,(X_j) = \frac{(t_k - t_{j'-1})^2}{12} + \Big(\frac{t_k + t_{j'-1}}{2} - x_j\Big)^2 \tag{47}$$

weil auch die Differenz zwischen der mittleren Bewertung einer Aussage und dem jeweiligen besten Schätzwert x_j einen Beitrag liefert. Der Unsicherheitskennwert u'_B charakterisiert somit nicht nur die Differenz zwischen minimaler und maximaler Bewertung in allen Aussagen des Beobachters, sondern zusätzlich die Differenz zwischen der mittleren Bewertung einer Aussage und dem jeweiligen besten Schätzwert des Merkmals in der Übung, zu der die Aussage gehört.

6.2.3 Angabe der Unsicherheitskennwerte

Die nach Abschnitt 6.2.2 berechneten Unsicherheitskennwerte zu JA, JB, JC und ST sind in den Tabellen A.4 bis A.7 in Anhang A aufgeführt.

In Spalte 1 dieser Tabellen steht die Merkmals-Nummer für die Spalten 2 bis 9 bzw. die Beobachter-Nummer für die Spalten 10 und 11. Die Merkmals-Nummer ist die nach Tabelle 5.2 bei JA, JB und JC bzw. die nach Tabelle 5.3 bei ST. Die Beobachter-Nummer ist genauer bei JA und JB die Nummer einer Beobachterperson und bei ST die Nummer eines bewertenden Teams, das als „Beobachter“ zu betrachten ist (Abschnitt 5.4.1). Die Tests sind ebenfalls als „Beobachter“ aufzufassen, auch für sie sind Unsicherheitskennwerte unterhalb der kurzen horizontalen Linien in den Spalten 10 und 11 aufgeführt. Den Tests der Übungen 1, 2 und 3 entsprechen bei JA die Beobachter-Nummern 5, 6 bzw. 7, bei JB die Beobachter-Nummern 13, 14 bzw. 15

und den Tests der Übungen 4 und 5 bei ST entsprechen die Beobachter-Nummern 8 bzw. 9. Bei JC in Tabelle A.6 sind Beobachter-Nummern nicht relevant (siehe unten).

Die Spalten 2 bis 8 sind den Übungen 1 bis 7 der Tabellen 5.2 bzw. 5.3 zugeordnet. In diesen Spalten stehen die Unsicherheitskennwerte u_U der Übungen k bezüglich der Merkmale i, d.h. die Matrix dieser Unsicherheitskennwerte. Bei JA, JB und JC entfällt Spalte 8, weil nur 6 Übungen durchgeführt wurden. Auch in ihrem Aufbau entsprechen die Spalten 1 bis 7 der Tabellen A.4 bis A.6 der Tabelle 5.2 und die Spalten 1 bis 8 von Tabelle A.7 entsprechen Tabelle 5.3. Zu einem Merkmal i, das in der Übung k nicht bewertet wurde, gibt es natürlich keinen zugehörigen Unsicherheitskennwert und damit auch keinen Tabelleneintrag und kein Matrixelement. Der größte Unsicherheitskennwert in der Matrix der u_U, d.h. aller Spalten 2 bis 8, ist fett gedruckt.

In Spalte 9 sind die Unsicherheitskennwerte u_M der Merkmale und in den Spalten 10 und 11 die beiden Unsicherheitskennwerte u_B bzw. u'_B der Beobachter aufgeführt. Fett gedruckt sind die in diesen Spalten jeweils größten Unsicherheitskennwerte. Zu JC sind Unsicherheitskennwerte der Beobachter in Tabelle A.6 nicht angegeben, weil sie identisch sind mit den entsprechenden Unsicherheitskennwerten zu JA und JB in den Tabellen A.4 und A.5, zu den Tests sind sie identisch mit den entsprechenden Werten von u_U in den Spalten 1 bis 3 der Tabelle A.6 (Abschnitt 6.2.4). (Es könnten sich zu JC zwar veränderte Unsicherheitskennwerte für Beobachter ergeben, die sowohl in JA als auch in JB tätig waren, doch ist es wegen der Anonymisierung der AC-Unterlagen nicht bekannt, ob es solche gibt.)

Auch die Unsicherheitskennwerte sind mit den charakteristischen Standardunsicherheiten u_1 und u_2 nach Abschnitt 4.5.2 ähnlich wie in Abschnitt 6.1.2 zu vergleichen. Bei einem Unsicherheitskennwert wird jedoch nur $L = 1$ „Merkmal“ bewertet. Damit ergeben sich nach den Gleichungen (43) und (44) für JA, JB und JC $u_1 = 0{,}289$ und $u_2 = 1{,}73$ und für ST $u_1 = 0{,}144$ und $u_2 = 1{,}30$. Die mit $*$ markierten Unsicherheitskennwerte in den Tabellen A.4 bis A.7 sind exakt gleich der charakteristischen, minimal möglichen Standardunsicherheit u_1, bewirkt allein durch die Skalenstufung.

6.2.4 Diskussion der Unsicherheitskennwerte

Der Vergleich aller in den Tabellen A.4 bis A.7 aufgeführten Unsicherheitskennwerte mit den vorstehend in Abschnitt 6.2.3 angegebenen charakteristischen Standardunsicherheiten u_1 und u_2 ergibt zunächst, dass keiner der Unsicherheitskennwerte kleiner als u_1 ist und dass alle kleiner als u_2 sind. Ersteres muss als Rechenprobe erwartet werden und letzteres zeigt, dass widersprüchliche Bewertungen keine merkliche Rolle

spielen. Die Informationsbedingung nach Gleichung (45), angewendet wie in Abschnitt 6.1.2 mit dem dort als sinnvoll erachteten $f = 1/2$, d.h. $fu_2 = 0{,}865$ bei JA, JB und JC bzw. mit $f = 3/4$, d.h. $fu_2 = 0{,}975$ bei ST, wird von fast allen Unsicherheitskennwerten erfüllt. Diese sind also akzeptabel. Lediglich die Unsicherheitskennwerte u_M des Merkmals 4 „Einfühlungsvermögen“ von JA und des Merkmals 3 „Lernmotivation, Belastbarkeit“ von ST werden als zu groß ausgewiesen (Spalte 9). Das bedeutet, dass diese Merkmale in den verschiedenen Übungen zu unterschiedlich bewertet wurden. In den Beobachterkonferenzen von JA wurde dies hinsichtlich des Merkmals 4 „Einfühlungsvermögen“ unabhängig von der Evaluierung an dieser Stelle auch so gesehen. Das ist der Grund, warum dieses Merkmal in Übung 5 „Präsentation“ von JB nicht mehr bewertet wurde (Tabelle 5.2). Mit mäßigem Erfolg, denn der zugehörige Unsicherheitskennwert u_M in Tabelle A.5 ist zwar etwas kleiner als der entsprechende in Tabelle A.4 und erfüllt die Informationsbedingung, ist aber immer noch der mit Abstand maximale Wert, der in Spalte 9 vorkommt.

Der vorstehende Absatz beantwortet die Fragen von Abschnitt 6.2.1 bereits weitgehend. Es werden aber noch die Unsicherheitskennwerte in den Tabellen A.4 bis A.7 miteinander verglichen, zunächst die Unsicherheitskennwerte u_U der Übungen bezüglich der Merkmale.

Unter diesen Unsicherheitskennwerten u_U befinden sich auch diejenigen der durchgeführten Tests in den Übungen 1 bis 3 von JA, JB und JC (Tabelle 5.2) sowie des Lerntests (Übung 4) von ST (Tabelle 5.3). Bei diesen Tests war nur jeweils ein einziges Merkmal zu bewerten. Das Testergebnis diente direkt als Bewertung dieses Merkmals und die minimale und maximale Bewertung waren gleich dem Testergebnis zu setzen (Abschnitt 4.3.3). Aus diesem Grund sind die zugehörigen Unsicherheitskennwerte u_U exakt gleich der charakteristischen, minimalen Standardunsicherheit u_1 und mit $*$ markiert. Ausnahme ist Übung 2 „Konzentrationstest d2“ von JA und JC (Spalte 3 der Tabellen A.4 und A.6). Hier ist der zugehörige Unsicherheitskennwert 0,35 bzw. 0,32 etwas größer als $u_1 = 0{,}289$, aber noch kleiner als die übrigen Unsicherheitskennwerte u_U in den Tabellen A.4 und A.6. Grund für diese Vergrößerung ist allein das fehlende Testergebnis für Teilnehmer 48 (Abschnitt 6.1.2). Diese Ausnahme zeigt, dass bei einem Test, dessen Ergebnis direkt als Bewertung eines einzigen Testmerkmals genommen wird, eine leichte Erhöhung des Unsicherheitskennwertes über u_1 hinaus auf die Unfähigkeit sehr weniger Teilnehmer hinweisen kann, den Test zu verstehen oder durchzustehen. Eine stärkere Erhöhung jedoch sollte Anlass geben, den Test kritisch zu prüfen, z.B. dahingehend, ob er verständlich genug formuliert ist.

Die übrigen Unsicherheitskennwerte u_U von JA, JB und JC in den Tabellen A.4 bis A.6 liegen zwischen 0,46 und 0,59 und unterscheiden sich nicht sehr stark in Anbetracht der weit darüber liegenden Grenze $fu_2 = 0{,}865$ der Informationsbedingung. Im Vergleich dazu streuen die übrigen Unsicherheitskennwerte u_U von ST in Tabelle A.7 (der Test in Übung 4 wurde schon besprochen) stärker und liegen auch systematisch höher zwischen 0,40 und 0,82, d.h. zumeist höher als der größte Wert 0,59 bei JA, JB und JC. Sie befinden sich aber noch genügend weit unterhalb der Grenze $fu_2 = 0{,}975$ der Informationsbedingung. Die größere Streuung und die systematische Vergrößerung in ST gegenüber den vorgenannten Werten in JA, JB und JC ist wieder eine Folge davon, dass in allen hier betrachteten Übungen von ST, auch bei dem kognitiven Test (Übung 5), Verhaltensanker bewertet wurden (Tabelle 5.3), nicht jedoch in JA, JB und JC. Die Ursache wurde bereits in Abschnitt 6.1.2 diskutiert.

Allerdings hat der hohe Unsicherheitskennwert $u_U = 0{,}82$ von Merkmal 2 „Unternehmerisches Denken“ der Postkorb-Übung (Übung 1) von ST noch eine zweite Ursache, die durch Nachschau in den Unterlagen von ST deutlich wurde: Der Verhaltensanker „Gegenleistung fordern“ dieses Merkmals blieb bei 18 der 25 Teilnehmer unbewertet. Offenbar ließ sich dieser Verhaltensanker nicht immer beobachten. Dahingehend bedarf die angewendete Postkorb-Übung einer Verbesserung.

In den Unsicherheitskennwerten u_M der Merkmale finden sich die besprochenen Charakteristika der Unsicherheitskennwerte u_U der Übungen bezüglich der Merkmale wieder. Wenn ein Merkmal i in nur einer einzigen Übung bewertet wird, gilt $u_M = u_U$. Anderenfalls ist u_M größer als jedes einzelne u_U aller Übungen, in denen das Merkmal i bewertet wurde, weil in u_M auch noch der Unsicherheitsbeitrag der Streuung der Bewertungen über die Übungen hinweg enthalten ist. Die Differenz zwischen u_M und u_U ist maximal bei Merkmal 4 „Einfühlungsvermögen“ von JA und JC, Merkmal 7 „Formale Kommunikationsfähigkeit“ von JB und Merkmal 3 „Lernmotivation, Belastbarkeit“ von ST und damit ist der Beitrag dieser Streuung jeweils am größten. Wie schon oben festgestellt, überschreitet u_M in zwei von diesen vier Fällen die Grenze der Informationsbedingung. Alle Unsicherheitskennwerte u_M von JA, JB und JC liegen zwischen 0,29 und 0,88, die von ST systematisch höher zwischen 0,73 und 1,01. Das kann wieder auf die größere Streuung der Bewertungen zurückgeführt werden, die damit verbunden ist, dass in ST die Verhaltensanker zu einem Merkmal bewertet wurden und nicht das Merkmal selbst wie in JA, JB und JC.

Jeder einzelne Unsicherheitskennwert u_U oder u_M von JC liegt, wie zu erwarten, zwischen den entsprechenden von JA und JB und letztere unterscheiden sich nur geringfügig. Daraus lässt sich schließen, dass diese Unsicherheitskennwerte nur wenig

von den Zufälligkeiten der Bewertung der einzelnen Teilnehmer beeinflusst sind und daher tatsächlich charakteristisch sind für die Übungen und Merkmale.

Es wurden die Unsicherheitskennwerte u_B und u'_B aller „Beobachter“ berechnet, d.h. nicht nur der Beobachterpersonen von JA und JB und der Teams von ST, sondern auch der Tests, die ja auch zu den Beobachtern zu zählen sind. In ST einigten sich mehrere Beobachterpersonen im Team gemeinsam auf jede Bewertung. Deshalb sind in ST keine Aussagen zum Bewertungsverhalten einzelner Beobachterpersonen möglich, sondern nur zu dem der 7 Teams. Die Unsicherheitskennwerte u_B und u'_B zu JC sind aus den in Abschnitt 6.2.3 und weiter unten genannten Gründen nicht aufgeführt. Bei den Unsicherheitskennwerten u_B und u'_B zeigt es sich, dass, wie nach Abschnitt 6.2.2 erwartet, bei allen Beobachtern $u'_B \geq u_B$ gilt.

Im Folgenden werden zunächst die Unsicherheitskennwerte u_B und u'_B der Beobachterpersonen und Teams betrachtet (oberhalb der kurzen horizontalen Linien in den Spalten 10 und 11 der Tabellen A.4, A.5 und A.7), wobei zuerst u'_B und dann u_B diskutiert wird. Anschließend werden die Unsicherheitskennwerte u_B und u'_B der Tests untersucht (unterhalb der kurzen horizontalen Linien in den Spalten 10 und 11 der Tabellen A.4, A.5 und A.7). Es wird daran erinnert, dass u_B das Verhalten eines Beobachters bezüglich nur seiner eigenen Aussagen charakterisiert, während u'_B sein Bewertungsverhalten auch in Bezug auf die Merkmalsergebnisse wiedergibt, die in den einzelnen Übungen aus allen zugehörigen Aussagen der Beobachter folgen. u'_B könnte man daher als ein Maß für die Treffsicherheit des Beobachters auffassen.

In JA sind die Werte u'_B der Beobachterpersonen nahezu gleich und von etwa gleicher Größe wie die Unsicherheitskennwerte u_U der Übungen 4 bis 6, in denen die Beobachter tätig waren. Das erscheint plausibel, weil u'_B und u_U nach Abschnitt 6.2.2 zwar aus unterschiedlich gemittelten Varianzen gebildet werden, aber unter Verwendung derselben Aussagen der Beobachter und derselben Schätzwerte x_j der Merkmale in den einzelnen Übungen. Gleiches gilt bezüglich u'_B für JB mit Ausnahme des Beobachters 11, dessen hoher Wert 0,75 von den jeweiligen, aus allen zugehörigen Aussagen der Beobachter folgenden Merkmalsergebnissen stark abweichende Bewertungen und damit eine geringere Treffsicherheit als bei den anderen Beobachtern erkennen lässt. Gleiches gilt auch für ST mit Ausnahme des kleinsten vorkommenden Wertes $u'_B =$ 0,44 für Team 6, der kleiner ist als fast alle Werte u_U der Übungen 1, 2, 3, 6 und 7, in denen die Teams tätig waren. Ob zufällig oder aus Erfahrung, die Bewertungen von Team 6 waren offenbar treffsicherer als die der anderen Teams. Das Team weist allerdings keine Besonderheit bei seinen Mitgliedern auf, z.B. hinsichtlich der Qualifikation für die Personalbeurteilung.

Auch die Unsicherheitskennwerte u_B in JA sind nahezu gleich. Sie sind deutlich größer als die charakteristische, minimale Standardunsicherheit $u_1 = 0{,}289$ von JA und JB. Das hat seine Ursache darin, dass in JA häufig jeweils eine minimale und maximale Bewertung abgegeben wurde. Auch in JB sind die Unsicherheitskennwerte u_B mit Ausnahme des Wertes für Beobachter 1 nahezu gleich, liegen aber auffällig dicht bei u_1. Zwei Werte stimmen sogar exakt damit überein. Der Ausnahmewert $u_B = 0{,}39$ für Beobachter 1 dagegen liegt, wie es auch bei den anderen Werten zu erwarten gewesen wäre, in seiner Größe nahe bei den Werten für u_B in JA. Es ist daher nicht der Ausnahmewert, der eine Besonderheit aufzuweisen scheint, sondern die übrigen Werte für u_B in JB sind es und bedürfen einer Erklärung. Ein Blick in die AC-Unterlagen zeigt aber, dass Beobachter 1 nur im AC 4 von JB mit nur vier Teilnehmern tätig war. Deshalb scheint der Ausnahmewert nur zufällig bei den Werten für u_B in JA zu liegen. Es gibt einen Unterschied zwischen JA und JB, der erklären kann, warum die Werte für u_B in JB so auffällig dicht bei u_1 liegen: In JB wurde im Vergleich zu JA nur spärlich von der Möglichkeit Gebrauch gemacht, jeweils eine minimale und maximale Bewertung abzugeben. Offenbar wurde bei der Beobachterschulung nicht deutlich genug auf diese Möglichkeit hingewiesen. Das ist denkbar, weil die Moderatorin von JA in JB nicht mehr tätig war.

Auch in ST sind die Unsicherheitskennwerte u_B zum Teil deutlich größer als die charakteristische, minimale Standardunsicherheit $u_1 = 0{,}144$ von ST und streuen mehr als in JA und JB. Das liegt hier jedoch nicht an der Abgabe jeweils einer minimalen und maximalen Bewertung, denn die Beobachter von ST waren gar nicht auf diese Möglichkeit hingewiesen wurden, obwohl sie sie trotzdem bei einigen seltenen Aussagen praktiziert haben. Das beruht auch n i c h t wieder auf einer relativ großen Streuung der Bewertungen der Verhaltensanker. Denn diese Steuung entfällt, weil zur Berechnung von u_B a l l e Aussagen des Beobachters so verschoben werden, dass die jeweilige mittlere Bewertung auf der Skalenmitte zu liegen kommt (Abschnitt 6.2.2). Die Ursache ist vielmehr, wie oben bei dem hohen Wert $u_U = 0{,}82$ von Merkmal 2 „Unternehmerisches Denken“ der Postkorb-Übung (Übung 1) von ST beschrieben, hauptsächlich darin zu suchen, dass bei der Mehrzahl der Teilnehmer die Bewertung des Verhaltensankers „Gegenleistung fordern“ zu diesem Merkmal 2 durch die Teams fehlt (außer bei Team 1), was sich auf den Unsicherheitkennwert u_B der Teams unterschiedlich auswirkt.

Die Unsicherheitskennwerte u_B und u'_B der Tests stehen unterhalb der kurzen horizontalen Linien in den Spalten 10 und 11 der Tabellen A.4, A.5 und A.7. Bei Tests, in denen nur ein einziges Merkmal und k e i n e Verhaltensanker bewertet werden und auch je Teilnehmer nur eine einzige Aussage vorliegt, gilt allgemein $u_B = u'_B = u_U$,

weil die Berechnungsverfahren in diesem Sonderfall übereinstimmen. u_U wurde schon oben diskutiert. Dies betrifft mit einer Ausnahme a l l e Tests der AC-Serien, auch JC. Deswegen sind u_B und u'_B für die Tests von JC nicht angegeben, sie sind gleich den entsprechenden Werten von u_U in den Spalten 1 bis 3 der Tabelle A.6. Die Ausnahme ist der kognitive Test von ST (Übung 5, Beobachter-Nummer 9), weil in diesem Test Verhaltensanker zu bewerten waren. Da aber nur ein einziges Merkmal bewertet wurde, gilt noch $u'_B = u_U$, wenn auch nicht mehr $u_B = u'_B$. Der Wert $u'_B = 0{,}77$ liegt im Rahmen der übrigen Werte u'_B von ST. Meist sind bei den Aussagen zu den Tests minimale und maximale Bewertung gleich, dann ist $u_B = u_1$ wie bei u_U und u_M.

Die Diskussion in diesem Abschnitt zeigt exemplarisch, wie die Unsicherheitskennwerte für die Evaluierung eines AC verwendet und interpretiert werden können.

6.3 Korrelationskoeffizienten

Üblich ist es, ein AC, das auf viele Teilnehmer angewendet worden ist, mit Hilfe von Korrelationskoeffizienten nach der konventionellen Statistik zu evaluieren (Kleinmann, 1997). Die Korrelationskoeffizienten bilden ein gutes Maß, um die Konstruktvalidität in einem AC kritisch zu untersuchen, insbesondere hinsichtlich der Konstruktvalidität einer Übung zu einem bestimmten Merkmal im Vergleich zu anderen Merkmalen und Übungen. Dies soll auch in diesem Abschnitt durchgeführt werden. Allerdings geht es hier weniger darum, die AC-Serien zu evaluieren, sondern die Unterschiede darzulegen, die zwischen der üblichen Anwendung der konventionellen Statistik einerseits und der Anwendung der Bayes'schen Statistik im Rahmen dieser Arbeit andererseits bestehen. Zu den Grundlagen der Korrelationskoeffizienten siehe Abschnitt 3.5.1.

Außerdem soll untersucht werden, ob die teilnehmerbezogene Korrelation (Abschnitt 3.5.2) vernachlässigt werden kann, was für die Berechnung von Unsicherheiten zusammengesetzter Merkmale nach den Abschnitten 3.4.3 und 4.4 Voraussetzung ist und nach Abschnitt 3.5.2 vermutet wird.

6.3.1 Berechnung und Angabe der Korrelationskoeffizienten

Für die Kennzeichnung und Berechnung der Korrelationskoeffizienten zur Evaluierung einer AC-Serie hinsichtlich der Konstruktvalidität wird einem Merkmal, das in einer bestimmten Übung bewertet wird, über alle Teilnehmer hinweg eine Zufallsvariable X_i zugeordnet. X_i gilt für das Merkmal n u r in dieser betrachteten Übung. Dabei bedeutet i die Kombination der Merkmals-Nummer und der Übungs-Nummer nach

Tabelle 5.2 für JA, JB und JC und nach Tabelle 5.3 für ST. Auf gleiche Weise wird demselben Merkmal in einer anderen Übung oder einem anderen Merkmal in derselben oder einer anderen Übung eine andere Zufallsvariable X_k zugeordnet. X_i und X_k entsprechen X bzw. Y in Abschnitt 3.5.1. Das Paar (i, k) kann somit dazu dienen, z.B. den Korrelationskoeffizienten $\varrho(X_i, X_k)$ zu kennzeichnen.

Für die Berechnung eines Korrelationskoeffizienten zur Konstruktvalidität zu einem Paar (i, k) wurden nun zu jedem Teilnehmer a l l e Paare von Aussagen zu X_i und X_k als die relevante Information herangezogen. Das gilt für die Berechnung sowohl nach der konventionellen als auch nach der Bayes'schen Statistik. Allerdings wurden bei der konventionellen Berechnung die minimalen und maximalen Bewertungen jeder Aussage, weil diese bisher unüblich sind, durch die mittlere Bewertung ersetzt. Dadurch werden die bei der konventionellen Berechnung des empirischen Korrelationskoeffizienten nach Gleichung (34) benötigten Werte x_j für X_i und y_j für X_k identisch mit den Erwartungswerten $\mathrm{E}\, X_i$ bzw. $\mathrm{E}\, X_k$ für den Teilnehmer j, die auch bei der Bayes'schen Berechnung der Varianzen und der Kovarianz in Gleichung (30) für den entsprechenden Korrelationskoeffizienten zu bilden sind. Weiteres zur Bayes'schen Berechnung siehe Abschnitt 3.5.2.

Alle berechneten Korrelationskoeffizienten zur Konstruktvalidität von JA, JB und JC sind in Tabelle A.8 des Anhangs A aufgeführt, die von ST in Tabelle A.9. Die Zeilen der Tabellen sind in Spalte 1 fortlaufend nummeriert. Jede Tabellenzeile gehört zu einem nach den Tabellen 5.2 oder 5.3 möglichen Paar (i, k) zweier Kombinationen i und k von Merkmals-/Übungs-Nummern. Die beiden Kombinationen sind in den Spalten 2 und 3 angegeben. Die Einträge in den Spalten 1 bis 3 dienen der Identifizierung der in der Tabellenzeile ab Spalte 5 folgenden Korrelationskoeffizienten. Alle Zeilen mit gleichem i stehen direkt untereinander und sind durch horizontale dünne Linien von den Zeilen mit anderen i getrennt.

In den Spalten 5, 7 und 9 stehen die nach der konventionellen Statistik für die Untersuchung der Konstruktvalidität nach Gleichung (34) entsprechend r_{xy} berechneten empirischen Korrelationskoeffizienten r_{ik} zu (i, k). In den Spalten 6, 8 und 10 folgen die nach der Bayes'schen Statistik und Gleichung (30) berechneten Korrelationskoeffizienten $\varrho_{ik} = \varrho(X_i, X_k)$ zu (i, k). Stimmen bei i und k entweder die Merkmals-Nummern oder die Übungs-Nummern überein, handelt es sich um Korrelationskoeffizienten zur konvergenten bzw. diskriminanten Konstruktvalidität. Das ist in Spalte 4 durch k bzw. d angezeigt. Die Korrelationskoeffizienten von JC stehen zwischen denen von JA und JB in Tabelle A.8, um deutlicher erkennbar werden zu lassen, ob bei JC

als Vereinigung von JA und JB zu erwartende „mittlere“ Werte auftreten. Die Korrelationskoeffizienten zu (k, i) sind nicht aufgeführt, da sie identisch sind mit denen zu (i, k). Die empirischen Korrelationskoeffizienten r_{ik} bilden die *Multitrait-Multimethod-Matrix* (MTMM-Matrix; Kleinmann, 1997).

Die in den Tabellen A.8 und A.9 zu JA, JB, JC und ST aufgelisteten Korrelationskoeffizienten zur Konstruktvalidität sind zur Übersicht auch in den Bildern A.1 bis A.4 in Anhang A als Histogramme dargestellt. Diese geben die Häufigkeitsverteilung der Korrelationskoeffizienten wieder. Die Abszisse ist in diesen Bildern in Intervalle der Länge 0,1 eingeteilt und als Ordinate ist die Anzahl der in diese Intervalle fallenden Korrelationskoeffizienten aufgetragen. Außerdem sind die Korrelationskoeffizienten zur konvergenten Konstruktvalidität durch • und die zur diskriminanten Konstruktvalidität durch ∘ angezeigt.

Bei sehr umfangreichem Datenmaterial und analogem Vorgehen unterscheiden die Ergebnisse der konventionellen und der Bayes'schen Statistik sich nicht wesentlich voneinander. Daher ist zu vermuten, dass auch die entsprechenden Korrelationskoeffizienten zur Konstruktvalidität sich nicht stark unterscheiden. Diese Vermutung kann noch wie folgt erweitert werden.

Unsicherheit auf dem Boden der Bayes'schen Statistik bedeutet, dass für ein Merkmal nicht nur ein bestimmter Schätzwert, sondern ein ganzer Bereich vernünftiger Schätzwerte infrage kommt (Abschnitt 3.3.2), was sich auf die Varianz im Vergleich zur konventionellen Statistik vergrößernd auswirkt und was man als „Verschmierung“ von Information auffassen kann. Geschieht diese Verschmierung bei jedem Merkmal u n a b h ä n g i g von anderen Merkmalen, hat dies keine Auswirkung auf die Kovarianz, da diese gleich null ist bei unabhängigen Zufallsvariablen. Das sollte dazu führen, dass bei der Berechnung eines konventionellen Korrelationskoeffizienten nach Gleichung (34) und eines Bayes'schen nach Gleichung (30) die Zähler, d.h. die Kovarianzen, sich praktisch nicht unterscheiden, der Nenner, genauer die Varianzen, jedoch im Bayes'schen Fall größer ist als der konventionell berechnete Nenner. Deshalb ist zu vermuten, dass der Quotient nach Gleichung (30) dem Betrag nach kleiner ist als der nach Gleichung (34), dass also die Korrelationskoeffizienten zur Konstruktvalidität nach der Bayes'schen Statistik dem Betrag nach kleiner ausfallen als die entsprechenden empirischen Korrelationskoeffizienten nach der konventionellen Statistik, obwohl sie sich nicht stark unterscheiden sollten. Diese erweiterte Vermutung ist zu prüfen.

Für die Kennzeichnung und Berechnung der teilnehmerbezogenen Korrelationskoeffizienten zur Untersuchung der Frage, ob die Korrelation bei der Berechnung von

Unsicherheiten vernachlässigt werden kann, wird einem Merkmal generell, d.h. über a l l e Übungen hinweg, eine Zufallsvariable X_j zugeordnet. Dabei bedeutet j die Merkmals-Nummer nach Tabelle 5.2 für JA, JB und JC und nach Tabelle 5.3 für ST. Auf gleiche Weise wird einem anderen Merkmal eine andere Zufallsvariable X_l zugeordnet. X_j und X_l entsprechen X bzw. Y in Abschnitt 3.5.1. Das Paar (j, l) kann somit dazu dienen, z.B. den teilnehmerbezogenen Korrelationskoeffizienten $\delta_{jl} = \varrho(X_j, X_l)$ zu kennzeichnen. Man beachte den Unterschied zu den Zuordnungen am Anfang dieses Abschnitts.

Für die Berechnung des teilnehmerbezogenen Korrelationskoeffizienten δ_{jl} zu einem Merkmal-Paar (j, l) wurden nun zu jedem Teilnehmer alle Aussagen zu X_j und X_l aus a l l e n Übungen als die relevante Information herangezogen. Die Berechnung erfolgte nach Gleichung (36). Alle Aussagen wurden reduziert, d.h., wie in den Abschnitten 3.5.2 und 6.2.1 beschrieben, auf den fiktiven „mittleren“ Teilnehmer bezogen und umgerechnet. Zu beachten ist, dass zur Kovarianz n u r Paare von Aussagen zu X_j und X_l d e s s e l b e n Beobachters als relevante Information beitragen, also k e i n e Paare von Aussagen unterschiedlicher Beobachter, weil letztere Aussagen als unabhängig voneinander entstanden angesehen werden können. Die Wurzeln aus den über alle Teilnehmer gemittelten Varianzen im Nenner des Quotienten in Gleichung (36) sind im vorliegenden Fall identisch mit den Unsicherheitskennwerten u_M (Abschnitt 6.2.2). Die zu mittelnden Varianzen und Kovarianzen sind jeweils mit den Erwartungswerten $\mathrm{E}\,X_j$ und $\mathrm{E}\,X_l$ der einzelnen Teilnehmer gebildet.

Alle berechneten teilnehmerbezogenen Korrelationskoeffizienten δ_{jl} von JA, JB, JC und ST sind in Tabelle A.10 des Anhangs A aufgeführt. Die Zeilen der Tabelle sind in Spalte 1 fortlaufend nummeriert. Jede Tabellenzeile gehört zu einem nach den Tabellen 5.2 oder 5.3 möglichen Paar (j, l) zweier Merkmale mit den Nummern j und l. Diese beiden Nummern des Merkmal-Paares sind in den Spalten 2 und 3 angegeben. Die Einträge in den Spalten 1 bis 3 dienen der Identifizierung der in den Tabellenzeilen 4 bis 7 folgenden Korrelationskoeffizienten. Alle Zeilen mit gleichem j stehen direkt untereinander und sind durch horizontale dünne Linien von den Zeilen mit anderen j getrennt. Die Korrelationskoeffizienten δ_{lj} sind nicht aufgeführt, da sie identisch sind mit δ_{jl}. Werte exakt gleich null sind in Tabelle A.10 ohne Kommastellen angegeben. Am Tabellenende sind noch die spaltenweise berechneten Mittelwerte der teilnehmerbezogenen Korrelationskoeffizienten und die empirischen Standardabweichungen dieser Mittelwerte (Abschnitt 3.5.1) aufgeführt. Die dem Betrag nach jeweils größten Korrelationskoeffizienten in den einzelnen Spalten sind fett gedruckt.

6.3.2 Diskussion der Korrelationskoeffizienten zur Konstruktvalidität

Es zeigt sich zunächst, dass in keiner Zeile der Tabellen A.8 und A.9 der Bayes'sche Korrelationskoeffizient zur Konstruktvalidität dem Betrag nach größer ist als der konventionelle. Die Differenz ist im Allgemeinen gering in JA, JB und JC, größer in ST. Die zu prüfende Vermutung nach Abschnitt 6.3.1 wird damit im Wesentlichen bestätigt. Dies zeigen auch die Verteilungen in den Bildern A.1 bis A.4. Wegen der dem Betrag nach kleineren Bayes'schen Korrelationskoeffizienten ist deren Verteilung etwas schmaler als die Verteilung der entsprechenden konventionellen Korrelationskoeffizienten. In ST ist dies deutlicher ausgeprägt als in JA, JB und JC, sonst aber ähneln sich die Verteilungen stark. Grund dafür ist, dass in ST generell die Unsicherheiten größer sind als in JA, JB und JC, was die Unsicherheitskennwerte u_U und u_M in den Tabellen A.4 bis A.7 erkennen lassen. Diese Unsicherheiten tragen zum Nenner des Quotienten in Gleichung (30) bei.

Die Bilder A.1 bis A.4 zeigen auch, dass alle Verteilungen der Korrelationskoeffizienten n i c h t um 0 zentriert sind. Positive Korrelationskoeffizienten überwiegen stark. Eine generelle positive Korrelation ist daher deutlich sichtbar. Im Idealfall sollten dagegen alle Korrelationskoeffizienten zur konvergenten Konstruktvalidität gleich 1 sein und alle übrigen gleich 0. Wenigstens sollten sie nahe 1 bzw. 0 liegen. Grund dafür ist, dass die beiden Zufallsvariablen X_i und X_k, auf die sich ein Korrelationskoeffizient bezieht, im konvergenten Fall d e m s e l b e n Merkmal, wenn auch in unterschiedlichen Übungen, zugeordnet sind, in anderen Fällen jedoch nicht. Deshalb sollten sie im konvergenten Fall streng voneinander abhängen, sonst aber unabhängig voneinander sein. Das lässt sich jedoch nicht erkennen, auch nicht ansatzweise. Im Gegenteil: während die Korrelationskoeffizienten zur konvergenten Konstruktvalidität zumeist mit positiven Werten recht zufällig verteilt sind, liefern ausgerechnet diejenigen zur diskriminanten Konstruktvalidität gerade die großen Werte, besonders in JA, JB und JC. Unterschiedliche Merkmale in derselben Übung sind also zu stark korreliert und dieselben Merkmale in unterschiedlichen Übungen zu schwach. Die Bayes'schen Korrelationskoeffizienten zur konvergenten Konstruktvalidität können zwar den Idealwert 1 wegen der Unsicherheit nicht ganz erreichen, sollten aber dennoch größer sein als diejenigen zur diskriminanten Konstruktvalidität. Mangelnde Konstruktvalidität wird hierin deutlich sichtbar, wie sie auch in ähnlichen anderen Studien immer wieder gefunden wird (z.B. Kleinmann, 1997; Russell, 1987; Scholz und Schuler, 1993).

Besonders klein sind die Werte der Korrelationskoeffizienten zur konvergenten Konstruktvalidität zu JA in Zeile 40 von Tabelle A.8 sowie zu ST in den Zeilen 106 und 111

von Tabelle A.9. Die von ST sind sogar negativ. Diese Werte weisen aufgrund zu geringer Korrelation auf mangelnde Konstruktvalidität hin. In JA betrifft dies das Merkmal 4 „Einfühlungsvermögen“ in den Übungen 4 und 5, das auch schon in den Abschnitten 5.3.2 und 6.2.4 aus anderen Gründen bemängelt wurde. In ST betrifft der Mangel die Übungen 2, 6 und 7 zum Merkmal 7 „Konfliktfähigkeit, Durchsetzungskraft“. Siehe hierzu auch die Tabellen 5.2 und 5.3. Zu starke Korrelation und damit ebenfalls mangelnde Konstruktvalidität zeigen die besonders großen Werte der Korrelationskoeffizienten zur diskriminanten Konstruktvalidität in Tabelle A.8 zu JA, JB und JC zwischen den folgenden Merkmalen an: zwischen Merkmal 4 „Einfühlungsvermögen“ und Merkmal 7 „Formale Kommunikationsfähigkeit“ in Übung 5 „Präsentation“ (Zeile 55, nur JA) sowie Merkmal 10 „Initiative, Motivation“ in Übung 6 „Selbstpräsentation, Interview“ (Zeile 69), zwischen Merkmal 5 „Konfliktfähigkeit“ und Merkmal 6 „Pädagogisches Geschick, Durchsetzungsfähigkeit“ in Übung 4 „Konfliktrollenspiel“ (Zeile 70) und zwischen Merkmal 8 „Reflexionsvermögen“ und Merkmal 10 „Initiative, Motivation“ in Übung 6 „Selbstpräsentation, Interview“ (Zeile 102). Bei ST sind hier in Tabelle A.9 zu nennen: zwischen Merkmal 6 „Teamfähigkeit“ und Merkmal 8 „Überzeugungskraft“ in den Übungen 6 und 7 „Rollenspiele“ (Zeilen 98 und 105) sowie zwischen Merkmal 7 „Konfliktfähigkeit, Durchsetzungskraft“ und Merkmal 8 „Überzeugungskraft“ in Übung 2 „Gruppendiskussion“ (Zeile 108). Dieser Absatz gilt für beide Statistiken.

Werden in jeder Zeile von Tabelle A.8 die drei sich entsprechenden empirischen Korrelationskoeffizienten r_{ik} nach der konventionellen Statistik von JA, JB und JC miteinander verglichen und auf gleiche Weise auch die drei Korrelationskoeffizienten ϱ_{ik} nach der Bayes'schen Statistik, so lässt sich feststellen, dass die Werte von JA, JB und JC im Allgemeinen stark differieren, obwohl sie bei sehr vielen Teilnehmern annähernd gleich sein sollten. Das weist auf noch vorhandene zufallsbedingte Einflüsse hin. Die Anzahl der Teilnehmer ist also noch n i c h t groß genug, um die Korrelationskoeffizienten zur Konstruktvalidität genau genug zu ermitteln. Immerhin liegt aber mit wenigen Ausnahmen der Wert von JC jeweils zwischen den Werten von JA und JB, was plausibel erscheint, weil JC alle Teilnehmer von JA und JB vereinigt. Die Ausnahmen (siehe Bemerkungen a und b in Spalte 4) können ihre Ursache nicht darin haben, dass bei der Vereinigung einige Bewertungen der Teilnehmer von JA entfernt werden mussten, um gleiche Verhältnisse für alle Teilnehmer zu schaffen (Abschnitt 5.3.4). Denn das davon allein betroffene Merkmal 4 „Einfühlungsvermögen“ in Übung 5 „Präsentation“ ist bei den Ausnahmen nicht beteiligt. Andererseits gibt es auch keinen zwingenden Grund dafür, dass diese Ausnahmen nicht vorkommen dürfen.

Die Diskussion in diesem Abschnitt zeigt, dass es keine grundsätzlichen und wesentlichen Unterschiede gibt zwischen den auf übliche Weise nach der konventionellen Statistik berechneten Korrelationskoeffizienten zur Konstruktvalidität und denen, die unter Berücksichtigung der Unsicherheit nach der Bayes'schen Statistik ermittelt wurden. Somit kann die Berücksichtigung der Unsicherheit über den bisherigen Stand der Forschung hinaus keine neuen Hinweise auf die Ursachen für die mangelnde Konstruktvalidität liefern. Die Ursachen liegen natürlich bei der Konstruktion der Merkmale und Übungen verborgen, sie lassen sich aber eher durch fachkritische Analyse als durch eine akribische Datenauswertung finden. Diese kann jene lediglich unterstützen. Die großen, offenbar zufallsbedingten Unterschiede zwischen den Werten sich entsprechender Korrelationskoeffizienten in den Zeilen von Tabelle A.8 bei JA, JB und JC lassen deutlich werden, dass für eine Analyse der Korrelationskoeffizienten zur Konstruktvalidität eine AC-Serie offenbar auf sehr viele, mindestens 100 Teilnehmer angewendet werden sollte.

Üblich ist es, eine auf viele Teilnehmer angewendete AC-Serie nicht nur durch eine konventionell-statistische Ermittlung der Korrelationskoeffizienten zur Konstruktvalidität, sondern auch mittels einer Faktorenanalyse dieser Korrelationskoeffizienten zu evaluieren (Kleinmann, 1997). In diesem Kapitel sind jedoch nicht die AC-Serien, sondern in erster Linie das in dieser Arbeit entwickelte Auswerteverfahren zu evaluieren. Außerdem unterscheiden sich, wie oben gezeigt, die konventionellen Korrelationskoeffizienten zur Konstruktvalidität von den entsprechenden Bayes'schen nicht sehr stark. Deshalb erübrigt sich hier eine Faktorenanalyse. Sie kann nur geringfügig andere Ergebnisse erbringen als eine Faktorenanalyse der konventionellen Korrelationskoeffizienten.

6.3.3 Diskussion der teilnehmerbezogenen Korrelationskoeffizienten

Die in Tabelle A.10 zu JA, JB, JC und ST aufgeführten teilnehmerbezogenen Korrelationskoeffizienten zu allen Merkmal-Paaren X_j und X_l wurden berechnet, um zu untersuchen, ob sie vernachlässigt werden können, wie es vermutet und für die Berechnung von Unsicherheiten verlangt wird.

Bei Betrachtung der teilnehmerbezogenen Korrelationskoeffizienten fällt sofort auf, dass sehr viele davon exakt gleich null sind. Das sind alle diejenigen, bei denen in JA, JB und JC die Merkmale 1, 2 oder 3 (Fall 1) und in ST die Merkmale 2, 4 oder 5 (Fall 2) beteiligt sind, sowie nur in JA außerdem diejenigen, bei denen $j = 5$ oder 6 und $l = 8$, 9 oder 10 sind (Fall 3). Diese drei Fälle lassen sich wie folgt begründen.

Die teilnehmerbezogene Korrelation zwischen zwei Merkmalen wird durch Aussagenpaare zu diesen beiden Merkmalen bewirkt, wobei jedes Paar aus Aussagen desselben

Beobachters (Person, Team oder Test) zu demselben Teilnehmer besteht (Abschnitt 3.5.2). Wenn es keine solchen Aussagenpaare gibt, ist der teilnehmerbezogene Korrelationskoeffizient zu den beiden Merkmalen exakt gleich null. Das ist der Fall, wenn die Bewertung dieser Merkmale völlig getrennt voneinander verläuft, alle Übungen und Beobachter des einen Merkmals sich von denen des anderen unterscheiden. Das ist die Begründung für die Fälle 1 und 3. Die Trennung ist im Fall 3 in JA strikt eingehalten, in JB durch andere Rotation der Beobachter jedoch nicht, was die teilnehmerbezogene Korrelation in JB und JC verursacht.

Der teilnehmerbezogene Korrelationskoeffizient ist auch dann exakt gleich null, wenn beide Merkmale bei jedem Teilnehmer nur von einem und demselben einzigen Beobachter bewertet werden (Abschnitt 3.2.2), wobei dieser Beobachter allerdings von Teilnehmer zu Teilnehmer wechseln kann. Dies erklärt Fall 2.

Man könnte aus den vielen auftretenden Werten gleich null schließen, dass dies die zu prüfende Vermutung bereits weitgehend bestätigt. Doch das wäre ein Trugschluss. Richtig ist jedoch der daraus folgende, fast triviale Ratschlag, ein AC immer so zu organisieren, dass nur die oben besprochenen drei Fälle vorliegen. Dann gibt es keine Korrelation in der gewonnenen Information. (Es sei hier daran erinnert, dass es in der Bayes'schen Statistik nur auf diese Information ankommt.) Jeder Teilnehmer sollte danach in jeder einzelnen Merkmal/Übung-Kombination von jeweils immer (möglichst mehreren) anderen Beobachtern bewertet werden. Das erfordert aber im Allgemeinen viele Beobachter im AC und wird meist die personellen Möglichkeiten übersteigen. Deshalb müssen auch andere Fälle untersucht werden. Denn die teilnehmerbezogene Korrelation ist bei der Berechnung der Unsicherheiten erst dann generell in allen AC vernachlässigbar, wenn die teilnehmerbezogenen Korrelationskoeffizienten in a l l e n möglichen Fällen genügend klein sind. Weiteres zum Thema Beobachterrotation siehe Kleinmann (1997) und Lammers (2000).

Es lässt sich nun feststellen, dass alle teilnehmerbezogenen Korrelationskoeffizienten δ_{ik} dem Betrag nach kleiner sind als 0,09 und im Allgemeinen auch weitaus kleiner als die Korrelationskoeffizienten zur Konstruktvalidität in den Tabellen A.8 und A.9. Es überwiegen positive Werte. Jeder Wert von JC liegt wie auch in Tabelle A.8 erwartungsgemäß mit einer Ausnahme (Zeile 30) zwischen den entsprechenden Werten von JA und JB. Deren Unterschied lässt den Einfluss des Zufalls noch erkennen. Die größten vorkommenden (fett gedruckten) Werte betreffen Merkmal-Paare, bei denen beide Merkmale in d e n s e l b e n Übungen bewertet wurden. In allen AC-Serien sind diese Übungen Rollenspiele, bei ST auch die Gruppendiskussion (Übung 2). Auch die entsprechenden Werte der Korrelationskoeffizienten zur Konstruktvalidität in

den Tabellen A.8 und A.9 sind sehr groß. Dies lässt auf eine möglicherweise durch einen Halo-Effekt erklärbare Abhängigkeit zwischen den Merkmalen schließen, in JA und JC zwischen Merkmal 4 „Einfühlungsvermögen“ und Merkmal 5 „Konfliktfähigkeit“, in JB zwischen Merkmal 6 „Pädagogisches Geschick, Durchsetzungsfähigkeit“ und Merkmal 7 „Formale Kommunikationsfähigkeit“ und in ST zwischen Merkmal 7 „Konfliktfähigkeit, Durchsetzungskraft“ und Merkmal 8 „Überzeugungskraft“.

Am Tabellenende sind die spaltenweise gebildeten Mittelwerte $\overline{\delta}$ der teilnehmerbezogenen Korrelationskoeffizienten und die empirischen Standardabweichungen dieser Mittelwerte (Abschnitt 3.5.1) angegeben. Es zeigt sich, dass jeder Mittelwert größer ist als die zugehörige etwa zweieinhalbfache empirische Standardabweichung des Mittelwertes. Daher ist die Abweichung der Mittelwerte von null signifikant auf einem Niveau von über 99 % und dementsprechend nicht zufällig. Eine geringe positive teilnehmerbezogene Korrelation muss somit als generell vorhanden festgestellt werden. Nun erhebt sich die Frage: Ist diese Korrelation gering genug, sodass sie bei der Berechnung der Unsicherheiten vernachlässigt werden kann? Die folgende Abschätzung soll diese Frage klären.

Für ein zusammengesetztes Merkmal Z als Funktion der Merkmale X_l in der Form $Z = F(X_1, \ldots, X_L)$ entsprechend Gleichung (21) lautet die allgemeine Formel nach DIN 1319-3 (1996) zur Fortpflanzung von Unsicherheiten unter Berücksichtigung der Korrelation

$$u^2(z) = \sum_{l=1}^{L} \Big(\frac{\partial F}{\partial X_l}\Big)^2 u^2(x_l) + 2 \sum_{j=1}^{L-1} \sum_{l=j+1}^{L} \frac{\partial F}{\partial X_j} \frac{\partial F}{\partial X_l}\, u(x_j, x_l) \tag{48}$$

Die Unsicherheiten der Gewichte G_l bzw. H_l sind hier außer Acht gelassen. $u(x_j, x_l)$ ist die Kovarianz der Merkmale X_j und X_l. Mit den teilnehmerbezogenen Korrelationskoeffizienten δ_{jl} für die Paare der Merkmale X_j und X_l gilt $u(x_j, x_l) = u(x_j)u(x_l)\delta_{jl}$.

In JA, JB und JC werden $L = 10$ und in ST werden $L = 8$ Merkmale X_l bewertet und mit gleichen Gewichten $h_l = 1/L$ zum Gesamtmerkmal Z eines Teilnehmers zusammengesetzt. Dann sind

$$Z = F(X_1, \ldots, X_L) = \frac{1}{L} \sum_{l=1}^{L} X_l \qquad \frac{\partial F}{\partial X_l} = \frac{1}{L} \tag{49}$$

Wird angenommen, dass alle Merkmale die gleiche Standardunsicherheit $u(x_l) = u$ besitzen, folgt $u^2(z) = u^2/L$ aus der ersten Summe in Gleichung (48) oder aus Gleichung (42), d.h. ohne Berücksichtigung der Korrelation und der globalen relativen

Standardunsicherheit u_r der Gewichte. Aus der vollständigen Gleichung (48) ergibt sich mit der Kovarianz $u(x_j, x_l) = u^2\delta_{jl}$ und dem Mittelwert $\overline{\delta}$ aller $L(L-1)/2$ teilnehmerbezogenen Korrelationskoeffizienten δ_{jl} dagegen der höhere Wert $u^2(z) = (u^2/L)(1 + (L-1)\overline{\delta})$, wodurch sich $u(z)$ um den Faktor $\sqrt{1 + (L-1)\overline{\delta}}$ vergrößert. Mit dem größten Mittelwert $\overline{\delta} = 0{,}0087$ von JA, JB und JC nach Tabelle A.10 ergibt sich 1,04 für diesen Faktor und somit eine Vergrößerung von $u(z)$ um 4 %, in ST mit $\overline{\delta} = 0{,}0057$ entsprechend eine solche um 2 %. Dies sind charakteristische Werte, die man unter Umständen durch geschickte Rotation der Beobachter im AC noch verkleinern kann. Trotzdem wird man in der Praxis eines AC mit einem Einfluss der Korrelation von bis etwa 10 % auf die Standardunsicherheit des Gesamtmerkmals eines Teilnehmers zu rechnen haben. Dies dürfte in den meisten Fällen hinnehmbar sein, sodass in diesem Rahmen die teilnehmerbezogene Korrelation wie verlangt vernachlässigt werden kann. Da der Trend des Einflusses der teilnehmerbezogenen Korrelation bei allen Teilnehmern in die gleiche Richtung geht, wirkt sich der Einfluss auf eine Auswahlentscheidung nach der Größe der Unsicherheit bei Teilnehmern mit nicht signifikant unterschiedlichen Ergebnissen kaum aus. Lediglich die Antwort auf die Frage, ob Teilnehmer-Ergebnisse sich signifikant unterscheiden, wird bei Vernachlässigung der Korrelation wegen der dann kleineren Standardunsicherheiten etwas öfter bejaht werden.

Für den Fall, dass die teilnehmerbezogene Korrelation trotz dieser Überlegungen als nicht vernachlässigbar erachtet wird, stehen Verfahren aus der physikalischen Messtechnik zu ihrer Berücksichtigung bereit [Uns]. Allerdings würde deren Anwendung einen erheblichen zusätzlichen mathematischen Aufwand erfordern. Dieser erscheint derzeit, da es überhaupt erst um die Einführung des Begriffs der Unsicherheit in die psychologische Praxis geht, noch als unvertretbar hoch.

7 Fazit und Ausblick

Zum Abschluss sollen die wesentlichen Aspekte und Ergebnisse der vorliegenden Arbeit kurz zusammengefasst und ein Ausblick auf mögliche sinnvolle Weiterentwicklungen gegeben werden.

7.1 Charakterisierung und Vorteile der Messunsicherheit

Zu den wichtigsten Aufgaben der Psychologie im Personalmanagement von Unternehmen und anderen Organisationen gehören Personalauswahl und Potenzialbeurteilung. Eine wirksame, hilfreiche und deshalb gebräuchliche Methode dafür ist das Assessment Center. Wie auch mit jedem anderen Urteils- und Entscheidungsprozess bleibt damit allerdings immer ein Risiko von Fehlentscheidungen aufgrund von Fehlschlüssen oder mangelnder Information verbunden. Der Einsatz der EDV kann helfen, das Risiko durch strikte Beachtung aller vorgegebenen Anforderungen und vorhandenen Information zu vermindern. Darüber hinaus bietet die Messunsicherheit die Möglichkeit, den bisher unbeachtet gebliebenen Mangel an Information und damit das Entscheidungsrisiko zu quantifizieren. Der Informationsmangel besteht immer, weil die für eine exakte Beurteilung nötige Information aus dem AC nicht vollständig gewonnen werden kann. Die Messunsicherheit drückt auch die Genauigkeit und Qualität der Beurteilung von Merkmalen im AC aus und schafft damit Vertrauen in die gewonnenen Ergebnisse.

Das ist der Grund, warum in dieser Arbeit der in der physikalischen Messtechnik wichtige Begriff der Messunsicherheit – kürzer auch Unsicherheit genannt – auf Assessment Center und allgemein auch auf dazu analoge Urteils- und Entscheidungsprozesse übertragen wird. Die Übertragung der Unsicherheit geschieht durch Identifizierung eines Merkmals im AC, einer zu bewertenden Eigenschaft eines Teilnehmers, mit einer physikalischen Messgröße und die Quantifizierung der Unsicherheit beruht auf dem einfachsten Ansatz, jede Aussage zu einem Merkmal nicht wie üblich vom Beobachter durch eine einzige Bewertung, sondern durch von ihm noch als sinnvoll erachtete minimale und maximale Bewertungen ausdrücken zu lassen. Dies sind die beiden wesentlichen neuen Ansätze, die den theoretischen und den praktischen Aspekt dieser Arbeit betreffen.

Der Begriff der Unsicherheit eines Merkmals fußt auf der Bayes'schen Statistik und charakterisiert den Mangel an Information über das Merkmal. Das ist das Besondere der Unsicherheit und erfordert eine neue Denkweise, die sich gänzlich von der des

„Messfehlers“ als Abweichung von der wahren Ausprägung des Merkmals unterscheidet. Zur Quantifizierung der Unsicherheit dient die Standardunsicherheit $u(x)$ zum besten Schätzwert x des Merkmals X, wobei x als Erwartungswert und $u(x)$ als Standardabweichung einer Verteilung berechnet werden, die mit Hilfe des Bernoulli'schen Prinzips und der im AC gewonnenen Information zum Merkmal aufzustellen ist. Diese Verteilung ist k e i n e Häufigkeitsverteilung zufällig auftretender Werte bei vielmals wiederholten Versuchen wie in der konventionellen Statistik, sondern repräsentiert die gerade vorhandene Information. Eine Standardunsicherheit $u(x)$, die größer ist als eine andere, besagt auch n i c h t , dass der zugehörige Schätzwert x „schlechter“ ist als der des anderen Merkmals, sondern nur, dass zu X weniger Information vorliegt und daraus z.B. ein höheres Entscheidungsrisiko folgt. Der Unsicherheitsbereich des Merkmals X mit den Grenzen $x - u(x)$ und $x + u(x)$ umfasst alle diejenigen Schätzwerte des Merkmals, die neben dem aus der vorliegenden Information folgenden besten Schätzwert x auch noch vernünftig sind. Es wird aber nicht behauptet, dass die gesuchte wahre Ausprägung des Merkmals sich tatsächlich in diesem Bereich befindet oder dass dies ein Bereich zufällig auftretender Werte bei wiederholten Versuchen ist.

Aus der Übertragung und Quantifizierung der Unsicherheit folgt das in Abschnitt 4.4 beschriebene Auswerteverfahren für ein AC, das auf der Grundlage eines Modells der Ermittlung bester Schätzwerte und der Berechnung der zugehörigen Standardunsicherheiten nicht nur für alle einzelnen Merkmale, sondern auch für aus solchen gewichtet zusammengesetzte Merkmale dienen kann, z.B. für das Gesamtmerkmal eines Teilnehmers. Dieses Auswerteverfahren vermeidet den methodischen Mangel üblicher Verfahren, Unsicherheit durch unzureichende Information und auch eine oft sinnvolle Gewichtung von Merkmalen unberücksichtigt zu lassen. Es erfordert zwar geringen Mehraufwand für die Datenerfassung, bietet aber dafür den Vorteil der Quantifizierung der Unsicherheit für jedes Merkmal und erlaubt EDV-Unterstützung, die weit über die Bildung von Mittelwerten und Standardabweichungen aufgrund der Streuung der Bewertungen mehrerer Beobachter hinausgeht. Das Verfahren gewährleistet darüber hinaus die strikte Beachtung aller vorgegebenen Anforderungen und vorliegenden Information.

Mit der Standardunsicherheit eines Merkmals werden in dem Auswerteverfahren mehrere Komponenten des Mangels an Information über das Merkmal erfasst: der Beitrag durch subjektive Einschätzung des Beobachters hinsichtlich der Genauigkeit seiner eigenen Bewertung, die Streuung der Bewertungen der Beobachter, sowie die Unsicherheiten aufgrund der Stufung der Bewertungsskala und der Festlegung der Gewichte. Nur die Streuung beruht auf statistisch erhobenen Daten und nur solche können mittels der konventionellen Statistik bearbeitet werden, die übrigen Komponenten nicht.

Das ist der wesentliche Grund für die Anwendung der Bayes'schen Statistik, in der allerdings Wahrscheinlichkeit wie eine Chance zum Wetten zu verstehen ist und nicht als relative Häufigkeit auftretender Werte bei vielmals wiederholten Versuchen wie in der konventionellen Statistik.

Ein weiterer Vorteil der Berechnung der Unsicherheit für jedes einzelne einfache oder zusammengesetzte Merkmal jedes Teilnehmers besteht darin, dass dadurch der kritische Vergleich mehrerer gleichartiger Merkmale sehr erleichtert wird. Überlappen sich die Unsicherheitsbereiche zweier solcher Merkmale von zwei Teilnehmern, so können deren Ergebnisse, d.h. die besten Schätzwerte der Merkmale, nicht als signifikant verschieden angesehen werden. In diesem Fall ist dem Merkmal mit der kleineren Unsicherheit das geringere Entscheidungsrisiko zuzusprechen. Auch ein Akzeptanzkriterium fußt auf ähnliche Weise auf der Unsicherheit: Ein Ergebnis zu einem Merkmal kann akzeptiert werden, wenn der Unsicherheitsbereich des Merkmals vollständig auf der besseren Seite des vorgegebenen Akzeptanzgrenzwertes liegt.

Auf der Messunsicherheit beruhen außerdem neue Validitätsmaße in Form von Unsicherheitskennwerten, die je nach zu untersuchender Fragestellung als geeignet „gemittelte“ Unsicherheiten gebildet werden können (Abschnitte 6.2, 7.2 und 7.3). Die Fragestellung ist dabei bezüglich des AC je nach Interesse weitgehend frei wählbar. Ein Unsicherheitskennwert charakterisiert die im AC erzielte Genauigkeit hinsichtlich der zugehörigen Frage.

7.2 Fazit der Evaluierung

In Abschnitt 6.1 wird das Auswerteverfahren anhand der auf viele Teilnehmer angewendeten AC-Serien JA, JB und ST aus der Praxis evaluiert. Dabei spielt die Unsicherheit eine herausragende Rolle. Es zeigt sich, dass die Teilnehmer-Ergebnisse und -Rangfolgen aus diesem Verfahren mit denen aus der üblichen Auswertung im Rahmen der Unsicherheit übereinstimmen, d.h. alle Unterschiede sind durch die Unsicherheit erklärlich. Dies entweder dadurch, dass die auf übliche Weise gewonnenen Ergebnisse in den Unsicherheitsbereich fallen und deshalb ebenso vernünftige Schätzwerte darstellen wie die besten Schätzwerte, das sind die Teilnehmer-Ergebnisse aus dem neuen Verfahren, oder aber anderenfalls dadurch, dass mittels einer ebenfalls auf der Unsicherheit beruhenden Informationsbedingung ein zu großer Informationsmangel festgestellt wird. Abweichungen zu den Teilnehmer-Rangfolgen aus den Beobachterkonferenzen lassen sich durch ein Merkmal erklären, das in den Konferenzen zusätzlich in Betracht gezogen, im AC aber nicht bewertet wurde, obwohl es hätte zur Bewertung vorgesehen werden müssen, weil es einer wichtigen Anforderung an die Teilnehmer entsprach.

Ebenfalls auf der Grundlage der Unsicherheit werden in Abschnitt 6.2 Unsicherheitskennwerte als geeignet „gemittelte“ Unsicherheiten definiert und berechnet, die bei der Evaluierung eines AC mit Vorteil als Validitätsmaße verwendet werden können, wenn dieses AC auf viele Teilnehmer angewendet wurde. Der Unsicherheitskennwert einer Übung bezüglich eines Merkmals gestattet es in Ergänzung zu anderen Methoden, die Konstruktvalidität der Übung für das Merkmal zu beurteilen, der Unsicherheitskennwert eines Merkmals sagt aus, wie genau das Merkmal im AC überhaupt ermittelt werden kann, und der Unsicherheitskennwert eines Beobachters charakterisiert dessen Bewertungsverhalten. Die Anwendung der Informationsbedingung auf einen Unsicherheitskennwert macht auch hier einen zu großen vorliegenden Informationsmangel erkennbar (Abschnitt 4.5.2).

Der Vergleich der Korrelationskoeffizienten zur Konstruktvalidität nach der üblichen konventionellen Statistik mit denen nach der Bayes'schen Statistik in Abschnitt 6.3.2 ergibt, dass diese dem Betrag nach etwas kleiner sind als jene. Dies lässt sich mit der neuen Art der Bewertung erklären. Jedoch weisen die sich entsprechenden Verteilungen der Korrelationskoeffizienten keine wesentlich unterschiedlichen Charakteristika auf. Wie auch in anderen Studien oft beobachtet, zeigt sich hierin wieder mangelnde Konstruktvalidität der untersuchten AC-Serien. Die Korrelationskoeffizienten zur konvergenten und diskriminanten Konstruktvalidität sind im Allgemeinen viel kleiner bzw. größer als erwartet werden muss. Neue Hinweise auf die Ursachen dafür über den bisherigen Stand der Forschung (Kleinmann, 1997) hinaus konnten nicht gefunden werden. Latente Merkmale wie beim Halo-Effekt (Abschnitt 6.3.3) können die Korrelationskoeffizienten wie beobachtet beeinflussen, müssen aber in jedem Einzelfall identifiziert werden, was in der Praxis kaum möglich ist. Daraus folgt, dass ein AC sehr sorgfältig auf der Grundlage einer detaillierten Anforderungsanalyse zweckentsprechend zu konstruieren ist. Es sollten, wenn möglich und sinnvoll, solche Merkmale zur Bewertung in den Übungen vorgesehen werden, deren Unabhängigkeit durch Evaluierung gesichert erscheint.

Die Berechnung der Unsicherheiten nach dem Auswerteverfahren dieser Arbeit setzt voraus, dass die von den Korrelationskoeffizienten zur Konstruktvalidität wohl zu unterscheidenden teilnehmerbezogenen Korrelationskoeffizienten gleich null sind oder nur zufallsbedingt von null abweichen und deshalb vernachlässigt werden können. Dass dies so ist, zeigt die Untersuchung auch dieser Korrelationskoeffizienten in Abschnitt 6.3.3. Allerdings ist doch eine geringe, signifikante positive teilnehmerbezogene Korrelation bei den betrachteten AC-Serien nicht übersehbar, die durchaus eine Vergrößerung der Unsicherheit des Gesamtmerkmals bei allen Teilnehmern um bis etwa 10 % bewirken kann, was zu tolerieren ist.

7.3 Zukünftige Untersuchungen und Weiterentwicklungen

Der erste Schritt bei der Weiterentwicklung des in dieser Arbeit vorgestellten Auswerteverfahrens mit Berücksichtigung der Unsicherheit ist naheliegend. Er besteht darin, das Verfahren nicht nur nachträglich, sondern in der Praxis eines laufenden AC wirklich anzuwenden, zuerst begleitend neben der üblichen Auswertung, später als integrierter Bestandteil des AC. Dazu ist der Einsatz der EDV zwingend erforderlich und deshalb auch die Erstellung einer umfassenden Software für diesen Zweck, d.h. eines Programmsystems, wie es in Abschnitt 4.6 skizziert ist. Dieses Programmsystem muss neben praxisgerechten Funktionen und moderner Gestaltung vor allem eine rationelle Dateneingabe und eine bequeme Handhabung bieten. Besonders wichtig ist auch eine anschauliche und überzeugende Visualisierung der Ergebnisse der Auswertung als Entscheidungshilfe für die Beobachterkonferenz. Das kann für die Akzeptanz des Verfahrens durch Veranstalter, Moderatoren und Beobachter eines AC sehr förderlich sein. Das Programmsystem sollte so konzipiert werden, dass es in ganz allgemeinen AC-analogen Urteils- und Entscheidungsprozessen angewendet werden kann. Beispiel dafür kann das im Anhang B beschriebene vorläufige Demonstrations- und Experimentierprogramm QWAHL für die Auswahl von Alternativen sein.

Durch Anwendung des Auswerteverfahrens in weiteren AC-Serien können zukünftig durch Bildung von Unsicherheitskennwerten nach dem in den Abschnitten 6.2.1 und 6.2.2 beschriebenen Prinzip auch Fragestellungen untersucht werden, zu deren Evaluierung das vorliegende Datenmaterial nicht ausgereicht hat. Dies könnte z.B. die Frage sein, ob Frauen und Männer sich in ihrem Bewertungsverhalten unterscheiden. Die zu dieser Frage gehörenden Unsicherheitskennwerte der Frauen bzw. Männer sind dann wie der Unsicherheitskennwert u_B eines Beobachters zu berechnen, nur sind dabei statt aller Aussagen dieses Beobachters nunmehr alle Aussagen aller Frauen bzw. Männer heranzuziehen. Die AC-Serien sollten in diesem Fall nicht nur viele Teilnehmer, sondern auch genügend viele Frauen und Männer als Beobachter umfassen. Es können auch Aussagen aus unterschiedlichen AC-Serien benutzt werden, wenn diese AC-Serien dieselbe Bewertungsskala aufweisen.

Eine weitere Frage von Interesse ist die nach der prognostischen Validität eines AC. Üblicherweise wird dazu der Korrelationkoeffizient zwischen einem geeigneten, im AC zum früheren Zeitpunkt 1 bewerteten Merkmal Y und einem zu einem späteren Zeitpunkt 2 erhobenen Merkmal Z wie ein Korrelationskoeffizient zur Konstruktvalidität nach der konventionellen Statistik berechnet. Als zu betrachtendes Merkmal Z könnte beispielsweise der Berufserfolg gewählt werden, gemessen an Karrierestufe, Einkommen und Vorgesetztenurteil als „Verhaltensanker“. Der Korrelationskoeffizient könnte auch

nach der Bayes'schen Statistik gebildet werden wie in Abschnitt 6.3.1 beschrieben. Ein großer Wert des Korrelationskoeffizienten weist dann auf vorhandene prognostische Validität hin.

Die prognostischen Validität eines AC hinsichtlich eines Merkmals Z, dessen Ausprägung zu einem späteren Zeitpunkt mittels der im AC gewonnenen Information vorherzusagen ist, lässt sich auch mit Hilfe von Unsicherheitskennwerten auf folgende Weise evaluieren, wenn zeitversetzte Bewertungen des Merkmals bei genügend vielen Teilnehmern auf derselben Bewertungsskala, aber nicht notwendig aus denselben Übungen vorliegen. Das zu betrachtende Merkmal kann variabel sein wie der Berufserfolg, kann aber auch als stabil erachtet werden wie die Lernfähigkeit. Die Evaluierung kann z.B. mit Hilfe der Bedingung nach Gleichung (8) geschehen. Zunächst werden für jeden Teilnehmer aus den Daten des im früheren Zeitpunkt 1 stattgefundenen AC das Vorhersageergebnis z_1 des Merkmals Z für den späteren Zeitpunkt 2 sowie die zugehörige Standardunsicherheit $u(z_1)$ berechnet, ebenso das Ergebnis z_2 des Merkmals und die zugehörige Standardunsicherheit $u(z_2)$ aus den im späteren Zeitpunkt 2 erhobenen Bewertungen. Anschließend werden aus diesen Standardunsicherheiten $u(z_1)$ und $u(z_2)$ aller Teilnehmer die Unsicherheitskennwerte $u_{M,1}$ bzw. $u_{M,2}$ des Merkmals nach Abschnitt 6.2.2 gebildet. Dann ist $U = \sqrt{u_{M,1}^2 + u_{M,2}^2}$ der Unsicherheitskennwert für die Merkmalsdifferenz zwischen Vorhersage und Ist zum Zeitpunkt 2. Diese Merkmalsdifferenz sollte gleich null sein, wenn das AC für das Merkmal prognostisch valide ist. Tatsächlich aber werden sich die Vorhersageergebnisse z_1 im Allgemeinen von den entsprechenden Ergebnissen z_2 der einzelnen Teilnehmer zum Zeitpunkt 2 unterscheiden. Dann lässt sich aus deren über alle Teilnehmer quadratisch gemittelten Differenz der Kennwert $D = \sqrt{\overline{(z_1 - z_2)^2}}$ gewinnen, der die Merkmalsdifferenz charakterisiert. Ist nun $D > kU$ analog Gleichung (8), so ist die charakteristische Differenz offenbar signifikant größer als ihr aufgrund der Unsicherheit zugebilligt werden kann. Prognostische Validität muss in diesem Fall verneint werden.

Nach Anwendung des Verfahrens in genügend vielen AC-Serien mit jeweils vielen Teilnehmern lässt sich auch die Frage genauer untersuchen, für die in dieser Arbeit keine zufriedenstellende Antwort gefunden werden konnte. Das ist die offen gebliebene Frage nach Festlegung eines vernünftigen generellen Wertes für den Faktor f in der Informationsbedingung nach Gleichung (45), mit deren Hilfe zu prüfen ist, ob die Unsicherheit eines Merkmals als ausreichend klein akzeptiert werden kann oder ob ein nicht mehr vertretbarer Informationsmangel vorliegt. In diesen weiteren Untersuchungen ist zu klären, ob sich generell für alle AC-Serien ein vernünftiger Wert für f festlegen lässt oder auch unterschiedliche Werte für verschiedene Typen von AC-Serien, z.B. sol-

chen, bei denen Merkmale entweder direkt wie in JA und JB oder indirekt über ihre Verhaltensanker wie in ST bewertet werden.

Ein späterer Schritt bei der Weiterentwicklung des Auswerteverfahrens betrifft die Berücksichtigung der teilnehmerbezogenen Korrelation bei der Berechnung der Unsicherheiten zusammengesetzter Merkmale. Hier ist zunächst genau zu untersuchen, ob nicht schon durch konsequente Weiterverfolgung des in Abschnitt 6.3.3 erwähnten Ratschlags, d.h. allein durch geschickte Rotation der Beobachter im AC, die Korrelation gänzlich vermieden werden kann, ohne dabei die gewinnbare Information zu schmälern, was wie die Korrelation eine Vergrößerung der Unsicherheit mit sich bringen würde. Falls das nicht möglich ist und wenn Abschätzungen zeigen, dass eine genauere Berechnung der Unsicherheiten trotz des dazu nötigen erheblichen Aufwandes überhaupt sinnvoll und erforderlich ist, kann die Korrelation wie bei entsprechenden, bereits verfügbaren Verfahren der physikalischen Messtechnik [Uns] in das Auswerteverfahren eingebracht werden.

Ein weiteres Anwendungsfeld für die Unsicherheit in der psychologischen Forschung bilden Probleme der Aufspürung latenter Merkmale z.B. als Ursache von Korrelationen. Beispiel eines solchen Problems ist es, das ursächliche Merkmal eines Halo-Effektes zu identifizieren. Zu den dabei benutzten Verfahren gehören die multiple Regression und die Faktorenanalyse. Auch bei solchen Problemen und Verfahren kann die Quantifizierung der Unsicherheit hilfreich sein. Für diesen Zweck stehen Verfahren der Ausgleichsrechnung unter Berücksichtigung der Messunsicherheit ebenfalls aus der physikalischen Messtechnik bereit.

7.4 Abschließende Betrachtungen

Die vorliegende Arbeit zeigt, dass die Berechnung der Unsicherheiten gerade für die Entscheidungsfindung in der Praxis wie auf dem hier untersuchten wichtigen Anwendungsgebiet der Personalauswahl, wo im Fall einer Fehlbesetzung hohe finanzielle und emotionale Kosten durch Kündigung oder Überforderung bei Unternehmen und Betroffenen entstehen können, über das Bisherige hinaus wirksame Unterstützung bietet.

Wenn alle Ergebnisse eines AC mit den zugehörigen Unsicherheiten beispielsweise durch das in Anhang B beschriebene Auswahlprogramm QWAHL bereits aufbereitet zur Beobachterkonferenz oder Entscheidung vorliegen, wird die zu treffende Entscheidung sehr anschaulich vorbereitet. Durch die Unsicherheit wird den Entscheidungsträgern verdeutlicht, ob Teilnehmer-Ergebnisse signifikant verschieden sind und wie hoch das

immer vorhandene, aber bisher in der Regel unberücksichtigt bleibende Entscheidungsrisiko ist. Sie können danach befinden, ob diese Unsicherheit toleriert werden kann oder gegebenenfalls weitere Untersuchungen nötig sind.

Das durchgeführte AC betreffend wird an Hand der Unsicherheitskennwerte aufgezeigt, wie gut einzelne Merkmale erfasst wurden, ob geeignete Übungen eingesetzt wurden und inwieweit sich Beobachter in ihrem Bewertungsverhalten unterscheiden. Ebenso wird deutlich, wie beispielsweise bei den AC-Serien JA und JB, ob zusätzliche implizite Merkmale zum Tragen kommen, die im AC selbst nicht erhoben wurden. Anhand der Ergebnisse kann dann geklärt werden, ob die Untersuchungen ergänzt werden müssen oder ob einzelne Übungen und Merkmale zur Entscheidung anders gewichtet Berücksichtigung finden sollten. Die Ergebnisse liefern auch Hinweise für Verbesserungen eines AC, das in Serie gehen soll.

Es zeigt sich aber auch, wie wichtig im Vorfeld die gründliche Konzipierung eines AC nach dem Stand der Wissenschaft ist. Dabei sollten die Anforderungen an das AC und dessen Durchführung nach DIN 33430 (2002) Beachtung finden. Insbesondere sind eine sorgfältige Anforderungsanalyse im Hinblick auf die Ziele des AC und die daraus resultierende zweckentsprechende Ableitung der zu bewertenden Merkmale und der Übungen sowie Festlegung der Gewichte unerlässlich. Eine noch so verfeinerte Datenauswertung kann diese Vorarbeit nicht ersetzen.

Literaturverzeichnis

[Uns] zitiert zusammenfassend die folgende Literatur zur Messunsicherheit: DIN 1319-3 (1996), DIN 1319-4 (1999), GUM (1993), Weise und Wöger (1992, 1999).

Arbeitskreis Assessment-Center (Hrsg.) (1992). *Standards der Assessment-Center-Technik.* München: Arbeitskreis Assessment-Center.

Brickenkamp, R. (1997). *Handbuch psychologischer und pädagogischer Tests* (2. Aufl.). Göttingen: Hogrefe.

Daumenlang, K. (1995). Intelligenztests. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollst. überarb. u. erw. Aufl., S. 540–548). Göttingen: Hogrefe.

DIN 1319-1 (1995). *Grundlagen der Meßtechnik – Teil 1: Grundbegriffe*. Berlin: Beuth.

DIN 1319-3 (1996). *Grundlagen der Meßtechnik – Teil 3: Auswertung von Messungen einer einzelnen Meßgröße, Meßunsicherheit*. Berlin: Beuth.

DIN 1319-4 (1999). *Grundlagen der Meßtechnik – Teil 4: Auswertung von Messungen, Meßunsicherheit*. Berlin: Beuth.

DIN 13303-1 (1982). *Stochastik – Wahrscheinlichkeitstheorie, Gemeinsame Grundbegriffe der mathematischen und der beschreibenden Statistik; Begriffe und Zeichen.* Berlin: Beuth.

DIN 13303-2 (1982). *Stochastik – Mathematische Statistik, Begriffe und Zeichen.* Berlin: Beuth.

DIN 33430 (2002). *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen.* Berlin: Beuth.

DIN 55350-21 (1982). *Begriffe der Qualitätssicherung und Statistik – Begriffe der Statistik, Zufallsgrößen und Wahrscheinlichkeitsverteilungen.* Berlin: Beuth.

DIN 55350-22 (1987). *Begriffe der Qualitätssicherung und Statistik – Begriffe der Statistik, Spezielle Wahrscheinlichkeitsverteilungen.* Berlin: Beuth.

Drees, H.B. (1994). *Untersuchung zur Validität eines Assessment Centers; Hinweise zur empirischen Überprüfung eines Assessment Centers unter besonderer Berücksichtigung unterschiedlicher Beobachtertrainings* (Dissertation). Aachen: Technische Hochschule, Philosophische Fakultät.

Eisenführ, F.; Weber, M. (1994). *Rationales Entscheiden* (2. Aufl.). Berlin: Springer.

Fisseni, H.-J.; Fennekels, G. (1995). *Das Assessment Center – Eine Einführung für Praktiker.* Göttingen: Verlag für Angewandte Psychologie.

Funke, U. (1991). Die Validität einer computergestützten Systemsimulation zur Diagnose von Problemlösekompetenz. In H. Schuler und U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 114–122). Stuttgart: Verlag für Angewandte Psychologie.

Funke, U. (1993). Computergestützte Eignungsdiagnostik mit komplexen dynamischen Szenarios. *Zeitschrift für Arbeits- und Organisationspsychologie 37,* 109–118.

Funke, U. (1995). Szenarien in der Eignungsdiagnostik und im Personaltraining. In B. Strauß und M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 145–218). Göttingen: Verlag für Angewandte Psychologie.

Gigerenzer, G.; Todd, P. M.; ABC Reasearch Group (1999). *Simple Heuristics That Make Us Smart.* New York: Oxford Universitiy Press.

Gigerenzer, G.; Selten, R. (Eds.) (2001). *Bounded rationality: The adaptive toolbox* (Dahlem Workshop Reports). o.O.: MIT Press.

GUM (1993). *Guide to the Expression of Uncertainty in Measurement.* Genf: International Organization for Standardization (ISO) – 1995: korrigierter Neudruck. – 1995: *Leitfaden zur Angabe der Unsicherheit beim Messen.* Berlin: Beuth. – 1999: Europäische Vornorm ENV 13005, Deutsche Fassung: DIN V ENV 13005, Berlin: Beuth.

Hasselmann, D.; Strauß, B. (1995). *Herausforderung Komplexität Baustein 2: Computersimulierte Problemlöseaufgaben für Management-Diagnostik und Training (Textilfabrik).* Hamburg: Windmühle.

Hossiep, R.; Paschen, M.; Mühlhaus, O. (2000). *Persönlichkeitstests im Personalmanagement. Grundlagen, Instrumente und Anwendungen.* Göttingen: Verlag für Angewandte Psychologie.

Jetter, W. (1996). *Effiziente Personalauswahl.* Stuttgart: Schäffer-Poeschel.

Jochmann, W. (1995). Entwicklung und Optimierung von Assessment-Bausteinen. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollst. überarb. u. erw. Aufl., S. 635–648). Göttingen: Hogrefe.

Jochmann, W. (Hrsg.) (1999). *Innovationen im Assessment Center. Entwicklungen, Alternativen und Einsatzmöglichkeiten im Change Management.* Stuttgart: Schäffer-Poeschel.

Jöreskog, K.G.; Sörbom, D. (1989). *LISREL 7 – A guide to the program and applications.* Chicago: SPSS Inc.

Jungermann, H. (1995). Entscheidungsanalytische Verfahren für Personalbeurteilung. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollst. überarb. u. erw. Aufl., S. 815–820). Göttingen: Hogrefe.

Jungermann, H.; Pfister, H.R.; Fischer, K. (1998). *Die Psychologie der Entscheidung: Eine Einführung.* Heidelberg, Berlin: Spektrum Akademischer Verlag.

Kepner-Tregoe (1971). *Rationales Management.* o.O.: Verlag Moderne Industrie.

Kleinmann, M. (1997). *Assessment Center: Stand der Forschung – Konsequenzen für die Praxis.* Göttingen: Verlag für Angewandte Psychologie.

Krumbach, P. (1999). Aktuelle Methoden und Einsatzgebiete der Anforderungsanalyse. In W. Jochmann (Hrsg.), *Innovationen im Assessment Center. Entwicklungen, Alternativen und Einsatzmöglichkeiten im Change Management* (S. 85–108). Stuttgart: Schäffer-Poeschel.

Lammers, F. (2000). Beobachterrotation und die Konstruktvalidität des Asssessment Centers. *Zeitschrift für Differentielle und Diagnostische Psychologie 21,* 270–278.

Lang–von Wins, T.; von Rosenstiel, L. (2000). Potentialfeststellungsverfahren. In M. Kleinmann und B. Strauß (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 73–99). Göttingen: Verlag für Angewandte Psychologie.

Laplace, P.S. (1820). *Théorie Analytique des Probabilités* (4. Aufl.). Paris: Courcier. – 1847: Nachdruck in *Œvres Completes de Laplace.* Band VII. Paris: Gauthier-Villars.

Lee, P.M. (1989). *Bayesian Statistics: An Introduction.* New York: Oxford University Press.

Lehment, T. (1999). Neuere Assessment-Center Bausteine. In W. Jochmann (Hrsg.), *Innovationen im Assessment Center. Entwicklungen, Alternativen und Einsatzmöglichkeiten im Change Management* (S. 109–127). Stuttgart: Schäffer-Poeschel.

Mattenklott, A. (1988). Diagnostische Urteilsbildung. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (S. 387–398). München, Weinheim: Psychologie Verlags Union.

Nabe, C.; Schmid, S.E. (1997). Kopf oder Zahl? Windows-Software hilft bei komplexen Entscheidungen. *ct Magazin für Computertechnik 5,* 256–269.

Nagel, S.S. (Ed.) (1993). *Computer-aided decision analysis: Theory and applications.* Westport: Quorum Books.

Niermeyer, R. (1999). Beobachterkompetenz. In W. Jochmann (Hrsg.), *Innovationen im Assessment Center. Entwicklungen, Alterativen und Einsatzmöglichkeiten im Change Management* (S. 156–179). Stuttgart: Schäffer-Poeschel.

Obermann, C. (1992). *Assessment Center: Entwicklung, Durchführung, Trends.* Wiesbaden: Gabler.

Obermann, C. (1995). Computergestützte Planspiele in der Mitarbeiterauswahl. In T. Geilhardt und T. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 401–410). Göttingen: Verlag für Angewandte Psychologie.

Russell, C.G. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal 30*, 817–826.

Sarges, W. (1995). Interviews. In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 475–489). Göttingen: Hogrefe.

Sarges, W. (Hrsg.) (1996). *Weiterentwicklung der Assessment Center Methode.* Göttingen: Verlag für Angewandte Psychologie.

Schmidt, F.L.; Hunter, J.E. (2000). Messbare Personenmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann und B. Strauß (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 15–42). Göttingen: Verlag für Angewandte Psychologie.

Scholz, G.; Schuler, H. (1993). Das Nomologische Netzwerk des Assessment Centers: Eine Metaanalyse. *Zeitschrift für Arbeits- und Organisationspsychologie 37*, 73–85.

Scholz, G. (1994). *Das Assessment Center: Konstruktvalidität und Dynamisierung.* Beiträge zur Organisationspsychologie, Band 14 (Hrsg.: Schuler, H.; Stehle, W.). Göttingen, Stuttgart: Verlag für Angewandte Psychologie.

Schubert, P. (1999). Feedback-Prozesse. In W. Jochmann (Hrsg.), *Innovationen im Assessment Center. Entwicklungen, Alternativen und Einsatzmöglichkeiten im Change Management* (S. 181–192). Stuttgart: Schäffer-Poeschel.

Strauß, B.; Kleinmann, M. (1996). Computersimulierte Szenarien im Assessment–Center. In W. Sarges (Hrsg.), *Weiterentwicklung der Assessment Center Methode* (S. 69–86). Göttingen: Verlag für Angewandte Psychologie.

Weise, K.; Wöger, W. (1992). *Eine Bayessche Theorie der Meßunsicherheit* (Bericht PTB-N-11). Braunschweig: Physikalisch-Technische Bundesanstalt. – 1993: A Bayesian theory of measurement uncertainty. *Measurement Science and Technology 4*, 1–11.

Weise, K.; Wöger, W. (1994). Comparison of two measurement results using the Bayesian theory of measurement uncertainty. *Measurement Science and Technology 5*, 879–882.

Weise, K.; Wöger, W. (1999). *Meßunsicherheit und Meßdatenauswertung*. Weinheim: Wiley-VCH.

Wickmann, D. (1990). *Bayes-Statistik.* Mathematische Texte, Band 4 (Hrsg.: Knocke, N.; Scheid, H.). Mannheim, Wien, Zürich: BI Wissenschaftsverlag, Bibliographisches Institut und F.A. Brockhaus.

Zimolong, B.; Rohrmann, B. (1988). Entscheidungshilfetechnologien. In D. Frey, C. Graf Hoyos und D. Stahlberg (Hrsg.), *Angewandte Psychologie* (S. 624–646). München, Weinheim: Psychologie Verlags Union.

Anhang A: Tabellen und Bilder zu Kapitel 6

Seite

Tabellen

Bilder

Tabelle A.1: Teilnehmer-Ergebnisse der AC-Serie JA

Bedeutung der Spalten:
1: laufende Teilnehmer-Nummer (T-Nr.)
2: AC-Nummer / Teilnehmer-Nummer in diesem AC (AC-T-Nr.)
3: Gesamtergebnis z_K: Mittelwert der Bewertungen der einzelnen Merkmale nach der Beobachterkonferenz
4: Gesamtergebnis z_T nach dem Verfahren dieser Arbeit
5: Standardunsicherheit $u(z_T)$ zu z_T nach Spalte 4, **fett:** maximaler vorkommender Wert
6: Rang R'_K nach der Beobachterkonferenz, +: positives Einstellungsvotum
7: Rang R_K nach Spalte 3
8: Rang R_T nach Spalte 4, +: positives Einstellungsvotum, weil $z_T + u(z_T) < 3$
9: Bemerkungen:
a und b: Ränge nach Spalten 6 und 7 bzw. 7 und 8 unterscheiden sich
c: z_K liegt nicht zwischen $z_T - u(z_T)$ und $z_T + u(z_T)$

1 lfd. T-Nr.	2 AC-Nr. / AC-T-Nr.	3 Erg. z_K	4 Erg. z_T	5 Uns. $u(z_T)$	6 Rang R'_K	7 Rang R_K	8 Rang R_T	9 Bem.
1	1 / 1	3.90	3.85	0.17	4	6	6	a
2	1 / 2	2.75	2.80	0.18	3	3	3 +	
3	1 / 3	3.15	3.11	0.15	5	4	4	a
4	1 / 4	2.30	2.29	0.20	2 +	2	2 +	
5	1 / 5	3.45	3.36	0.22	6	5	5	a
6	1 / 6	2.25	2.25	0.14	1 +	1	1 +	
7	2 / 1	3.75	3.70	0.17	5	5	5	
8	2 / 2	2.00	2.01	0.14	1 +	1	1 +	
9	2 / 3	3.00	2.94	0.22	3	2	2	a
10	2 / 4	3.10	3.00	0.19	4	3	3	a
11	2 / 5	3.35	3.35	0.16	2	4	4	a
12	3 / 1	3.55	3.54	0.13	3	3	4	b
13	3 / 2	3.70	3.68	0.17	2	5	5	a
14	3 / 3	3.40	3.41	0.21	5	2	2	a
15	3 / 4	3.60	3.48	0.19	4	4	3	b
16	3 / 5	2.55	2.60	0.17	1 +	1	1 +	
17	3 / 6	4.20	4.12	0.20	6	6	6	
18	4 / 1	3.50	3.50	0.22	4	4	4	
19	4 / 2	3.05	3.05	0.16	2 +	3	3	a
20	4 / 3	2.65	2.69	0.17	1 +	1	1 +	
21	4 / 4	2.80	2.81	0.16	3	2	2 +	a
22	5 / 1	2.75	2.73	0.22	1 +	1	1 +	
23	5 / 2	3.75	3.77	0.19	5	4	4	a
24	5 / 3	3.03	3.09	0.22	2	2	2	
25	5 / 4	3.90	3.88	0.16	3	5	5	a
26	5 / 5	3.30	3.28	0.16	4	3	3	a
27	6 / 1	2.93	2.89	0.22	2	2	2	
28	6 / 2	2.59	2.63	0.18	1 +	1	1 +	
29	6 / 3	3.65	3.57	0.19	3	3	3	

30	7 / 1	3.28	3.25	0.20	5	5	5	
31	7 / 2	3.60	3.59	0.18	4	7	7	a
32	7 / 3	2.70	2.73	0.14	1 +	2	2 +	a
33	7 / 4	2.90	2.89	0.22	7	3	3	a
34	7 / 5	3.56	3.48	0.18	6	6	6	
35	7 / 6	3.15	3.11	0.18	3	4	4	a
36	7 / 7	2.50	2.40	0.15	2 +	1	1 +	a
37	8 / 1	2.60	2.59	0.16	3	3	3 +	
38	8 / 2	3.60	3.61	0.15	5	5	5	
39	8 / 3	2.80	2.87	0.16	4	4	4	
40	8 / 4	1.60	1.64	0.16	1 +	1	1 +	
41	8 / 5	1.70	1.69	0.19	2 +	2	2 +	
42	9 / 1	3.22	3.23	0.25	1 +	3	3	a
43	9 / 2	2.40	2.39	0.19	2	1	1 +	a
44	9 / 3	3.65	3.68	0.15	6	5	5	a
45	9 / 4	2.85	2.75	0.18	3	2	2 +	a
46	9 / 5	3.75	3.72	0.16	5	6	6	a
47	9 / 6	3.50	3.46	0.16	4	4	4	
48	10 / 1	4.50	4.12	**0.36**	2	2	2	c
49	10 / 2	3.38	3.34	0.20	1	1	1	
50	11 / 1	2.20	2.17	0.16	2 +	1	1 +	a
51	11 / 2	2.85	2.85	0.19	3	3	3	
52	11 / 3	3.25	3.23	0.16	4	4	4	
53	11 / 4	2.35	2.36	0.15	1 +	2	2 +	a
54	11 / 5	3.75	3.70	0.17	5	5	5	
55	11 / 6	3.83	3.80	0.24	6	6	6	
56	11 / 7	3.86	3.92	0.20	8	8	8	
57	11 / 8	3.85	3.86	0.18	7	7	7	
58	12 / 1	2.24	2.24	0.24	1 +	1	1 +	
59	12 / 2	3.95	3.95	0.21	7	6	6	a
60	12 / 3	2.60	2.58	0.23	2 +	2	2 +	
61	12 / 4	3.05	2.98	0.24	3 +	3	3	
62	12 / 5	3.55	3.53	0.17	4	5	5	a
63	12 / 6	3.39	3.40	0.21	5	4	4	a
64	12 / 7	4.02	4.01	0.22	6	7	7	a
65	13 / 1	3.50	3.40	0.21	6	4 5	4	a
66	13 / 2	2.30	2.34	0.20	2	1	1 +	a
67	13 / 3	2.95	2.93	0.15	3	2	2	a
68	13 / 4	3.90	3.80	0.15	5	8	8	a
69	13 / 5	3.75	3.71	0.19	4	7	6 7	a
70	13 / 6	3.25	3.16	0.23	1	3	3	a
71	13 / 7	3.65	3.71	0.19	8	6	6 7	a
72	13 / 8	3.50	3.42	0.17	7	4 5	5	a
73	14 / 1	3.25	3.23	0.18	3	2 3	2	
74	14 / 2	3.65	3.60	0.15	6	5	5	a
75	14 / 3	3.10	3.13	0.18	2	1	1	a
76	14 / 4	3.25	3.27	0.13	1 +	2 3	3	a
77	14 / 5	3.79	3.82	0.22	5	6	6	a
78	14 / 6	3.30	3.28	0.18	4	4	4	

Tabelle A.2: Teilnehmer-Ergebnisse der AC-Serie JB

Legende siehe Tabelle A.1
Spalte 6, T-Nr. 45 bis 48: Ränge wurden im AC nicht vergeben

1 lfd. T-Nr.	2 AC-Nr. / AC-T-Nr.	3 Erg. z_K	4 Erg. z_T	5 Uns. $u(z_T)$	6 Rang R'_K	7 Rang R_K	8 Rang R_T	9 Bem.
1	1 / 1	2.50	2.48	0.14	1 +	1	1 +	
2	1 / 2	3.15	3.14	0.13	2	2	2	
3	1 / 3	4.10	4.11	0.12	3	3	3	
4	2 / 1	2.40	2.51	0.21	1	1	1 +	
5	2 / 2	2.70	2.71	0.13	2	2	2 +	
6	2 / 3	3.15	3.08	0.14	3	4	4	a
7	2 / 4	2.95	2.95	0.21	4	3	3	a
8	3 / 1	4.35	4.41	0.17	5	7	7	a
9	3 / 2	2.55	2.60	0.16	1 +	2	2 +	a
10	3 / 3	2.45	2.37	0.15	2 +	1	1 +	a
11	3 / 4	3.60	3.54	0.17	3	4	4	a
12	3 / 5	4.30	4.30	0.15	7	6	6	a
13	3 / 6	3.25	3.22	0.16	4	3	3	a
14	3 / 7	4.00	4.15	0.19	6	5	5	a
15	4 / 1	3.05	3.13	0.15	1	1	1	
16	4 / 2	4.90	4.78	0.14	4	4	4	
17	4 / 3	3.65	3.78	0.25	3	3	3	
18	4 / 4	3.50	3.50	0.22	2	2	2	
19	5 / 1	3.70	3.73	**0.31**	5	4	4	a
20	5 / 2	4.35	4.38	0.13	9	10	10	a
21	5 / 3	3.85	3.83	0.25	4 +	5	5	a
22	5 / 4	3.30	3.35	0.13	2 +	2	2	
23	5 / 5	4.25	4.25	0.14	10	9	9	a
24	5 / 6	3.95	3.98	0.13	6	6 7	7	
25	5 / 7	4.00	3.91	0.17	8	8	6	b
26	5 / 8	3.95	3.98	0.15	7	6 7	8	b
27	5 / 9	3.35	3.35	0.16	3 +	3	3	
28	5 / 10	2.25	2.24	0.14	1 +	1	1 +	
29	6 / 1	4.05	4.08	0.22	4	4	4	
30	6 / 2	3.85	3.80	0.21	2	3	3	a
31	6 / 3	3.75	3.65	0.21	3	2	2	a
32	6 / 4	4.30	4.47	0.17	5	5	5	
33	6 / 5	4.40	4.55	0.26	6	6	6	
34	6 / 6	2.90	2.93	0.18	1 +	1	1	
35	7 / 1	2.90	2.89	0.18	2	2	2	
36	7 / 2	3.65	3.63	0.19	3	3	3	
37	7 / 3	4.80	4.80	0.18	4	6	6	a
38	7 / 4	1.95	1.98	0.16	1 +	1	1 +	
39	7 / 5	4.35	4.33	0.18	5	5	5	
40	7 / 6	4.30	4.30	0.19	6	4	4	a
41	8 / 1	4.35	4.37	0.18	3	4	4	a
42	8 / 2	4.10	4.13	0.19	2	1 2 3	2	
43	8 / 3	4.10	4.12	0.15	4	1 2 3	1	a

44	8 / 4	4.10	4.15	0.14	1	1 2 3	3	
45	9 / 1	3.27	3.36	0.17	–	4	4	
46	9 / 2	2.53	2.50	0.17	–	1	1 +	
47	9 / 3	3.09	3.09	0.21	–	2	2	
48	9 / 4	3.26	3.29	0.22	–	3	3	
49	10 / 1	3.12	3.13	0.18	3	3	3	
50	10 / 2	4.87	4.83	0.22	6	6	6	
51	10 / 3	2.79	2.85	0.21	2 +	1	1	a
52	10 / 4	3.76	3.79	0.18	5	5	5	
53	10 / 5	3.75	3.75	0.17	4	4	4	
54	10 / 6	2.98	2.99	0.17	1 +	2	2	a

Tabelle A.3: Teilnehmer-Ergebnisse der AC-Serie ST

Legende siehe Tabelle A.1
Spalte 6 entfällt, weil Werte mit denen in Spalte 7 identisch
Spalte 7: +: positives Weiterbildungsvotum der Beobachterkonferenz
Spalte 8: +: positives Votum, weil $z_T - u(z_T) > 3$

1 lfd. T-Nr.	2 AC-Nr. / AC-T-Nr.	3 Erg. z_K	4 Erg. z_T	5 Uns. $u(z_T)$	6	7 Rang R_K	8 Rang R_T	9 Bem.
1	1 / 1	3.44	3.35	0.38	–	3 +	3	
2	1 / 2	3.43	3.30	0.28	–	4 +	4 +	
3	1 / 3	2.88	2.97	0.33	–	7 +	7	
4	1 / 4	3.14	3.08	0.27	–	6 +	6	
5	1 / 5	3.60	3.52	0.31	–	2 +	2 +	
6	1 / 6	3.71	3.69	0.30	–	1 +	1 +	
7	1 / 7	3.21	3.23	0.28	–	5 +	5	
8	2 / 1	3.26	3.12	0.33	–	6	6	
9	2 / 2	3.70	3.68	0.25	–	3 +	3 +	
10	2 / 3	2.48	2.50	0.34	–	7	7	
11	2 / 4	3.88	3.84	0.24	–	2 +	1 +	b
12	2 / 5	3.96	3.77	0.28	–	1 +	2 +	b
13	2 / 6	3.68	3.53	0.31	–	4 +	4 +	
14	2 / 7	3.54	3.49	0.22	–	5 +	5 +	
15	3 / 1	3.63	3.52	0.26	–	3 +	3 +	
16	3 / 2	2.80	2.83	0.27	–	10	10	
17	3 / 3	3.69	3.63	0.21	–	2 +	2 +	
18	3 / 4	2.73	2.70	**0.39**	–	11	11	
19	3 / 5	3.76	3.71	0.24	–	1 +	1 +	
20	3 / 6	3.45	3.36	0.28	–	5 +	5 +	
21	3 / 7	3.04	3.08	0.21	–	7 +	6	b
22	3 / 8	3.06	3.01	0.29	–	6 +	7	b
23	3 / 9	3.51	3.49	0.27	–	4 +	4 +	
24	3 / 10	2.84	2.86	0.27	–	9	9	
25	3 / 11	2.95	2.97	0.34	–	8 +	8	

Tabelle A.4: Unsicherheitskennwerte der AC-Serie JA

Bedeutung der Spalten:
1: Merkmals-Nummer (M-Nr.) i für Spalten 2 bis 9 oder Beobachter-Nummer (B-Nr.) i für Spalten 10 und 11
2 bis 8: Unsicherheitskennwerte u_U der Übungen Uk bezüglich der Merkmale Mi nach Tabelle 5.2. Zeile 3 der Kopfleiste: Übungs-Nummer (U-Nr.) k
9: Unsicherheitskennwerte u_M der Merkmale
10 und 11: Unsicherheitskennwerte u_B und u'_B der Beobachter (bzw. der Tests unterhalb der kurzen waagerechten Linien)
Bemerkungen:
*: minimaler Wert exakt gleich Unsicherheit aufgrund der Skalenstufung
fett: maximale vorkommende Werte in den Spalten 2 bis 8 sowie in den Spalten 9, 10 bzw. 11

1	2	3	4	5	6	7	8	9	10	11
M-/B-	U-	—	Unsicherheitskennwerte u_U			—				
Nr. i	Nr. $k=1$	2	3	4	5	6		u_M	u_B	u'_B
1	0.29*	–	–	–	–	–	–	0.29*	0.41	0.51
2	–	0.35	–	–	–	–	–	0.35	**0.43**	**0.53**
3	–	–	0.29*	–	–	–	–	0.29*	0.40	0.51
4	–	–	–	0.50	0.50	0.54	–	**0.88**	0.42	0.53
5	–	–	–	0.53	–	–	–	0.53	0.29*	0.29*
6	–	–	–	0.53	–	–	–	0.53	0.35	0.35
7	–	–	–	0.52	0.50	–	–	0.76	0.29*	0.29*
8	–	–	–	–	0.52	0.53	–	0.72	–	–
9	–	–	–	–	0.50	**0.54**	–	0.73	–	–
10	–	–	–	–	–	0.54	–	0.54	–	–

Tabelle A.5: Unsicherheitskennwerte der AC-Serie JB

Legende siehe Tabelle A.4

1	2	3	4	5	6	7	8	9	10	11
M-/B-	U-	—	Unsicherheitskennwerte u_U			—				
Nr. i	Nr. $k=1$	2	3	4	5	6		u_M	u_B	u'_B
1	0.29*	–	–	–	–	–	–	0.29*	**0.39**	0.57
2	–	0.29*	–	–	–	–	–	0.29*	0.30	0.58
3	–	–	0.29*	–	–	–	–	0.29*	0.30	0.49
4	–	–	–	**0.59**	–	0.55	–	**0.82**	0.29	0.53
5	–	–	–	0.49	–	–	–	0.49	0.31	0.52
6	–	–	–	0.51	–	–	–	0.51	0.31	0.45
7	–	–	–	0.50	0.49	–	–	0.76	0.31	0.34
8	–	–	–	–	0.46	0.53	–	0.70	0.33	0.36
9	–	–	–	–	0.49	0.51	–	0.70	0.29*	0.45
10	–	–	–	–	–	0.55	–	0.55	0.29	0.45
11	–	–	–	–	–	–	–	–	0.30	**0.75**
12	–	–	–	–	–	–	–	–	0.29*	0.51
13	–	–	–	–	–	–	–	–	0.29*	0.29*
14	–	–	–	–	–	–	–	–	0.29*	0.29*
15	–	–	–	–	–	–	–	–	0.29*	0.29*

Tabelle A.6: Unsicherheitskennwerte der AC-Serie JC

Legende siehe Tabelle A.4, Spalten 10 und 11 entfallen

1	2	3	4	5	6	7	8	9
M-	U-	—	Unsicherheitskennwerte u_U			—		
Nr. i	Nr. $k=1$	2	3	4	5	6		u_M
1	0.29*	–	–	–	–	–	–	0.29*
2	–	0.32	–	–	–	–	–	0.32
3	–	–	0.29*	–	–	–	–	0.29*
4	–	–	–	0.54	–	0.54	–	**0.85**
5	–	–	–	0.52	–	–	–	0.52
6	–	–	–	0.52	–	–	–	0.52
7	–	–	–	0.51	0.49	–	–	0.76
8	–	–	–	–	0.50	0.53	–	0.71
9	–	–	–	–	0.49	0.53	–	0.72
10	–	–	–	–	–	**0.54**	–	0.54

Tabelle A.7: Unsicherheitskennwerte der AC-Serie ST

Legende siehe Tabelle A.4, Spalten 2 bis 8 nach Tabelle 5.3

1	2	3	4	5	6	7	8	9	10	11
M-/B-	U-	—	Unsicherheitskennwerte u_U			—				
Nr. i	Nr. $k=1$	2	3	4	5	6	7	u_M	u_B	u'_B
1	0.40	–	–	–	0.77	–	–	0.82	0.30	0.76
2	**0.82**	–	–	–	–	–	–	0.82	0.32	0.73
3	–	–	0.72	0.14*	–	–	–	**1.01**	**0.40**	0.78
4	–	–	0.74	–	–	–	–	0.74	0.19	0.59
5	–	–	0.59	–	–	–	–	0.59	0.29	0.66
6	–	0.71	–	–	–	0.82	0.75	0.88	0.21	0.44
7	–	0.73	–	–	–	0.71	0.68	0.90	0.36	**0.80**
8	–	0.47	–	–	–	0.71	0.54	0.73	0.14*	0.14*
9	–	–	–	–	–	–	–	–	0.14*	0.77

Tabelle A.8: Korrelationskoeffizienten zur Konstruktvalidität der AC-Serien JA, JB und JC

Bedeutung der Spalten:
1: laufende Zeilen-Nummer
2 und 3: Merkmals-Nummer (M-Nr.) / Übungs-Nummer (U-Nr.) des Partners i bzw. des Partners k des Paares (i, k) (siehe Tabelle 5.2)
4: Konstruktvalidität (Val., k konvergente, d diskriminante) und Bemerkungen (a, b siehe unten)

Korrelationskoeffizienten (K.) zur Konstruktvalidität zum Paar (i, k):
5, 7, 9: empirische K. r_{ik} nach der konventionellen Statistik
Bemerkung a in Spalte 4: Wert in Spalte 7 liegt nicht zwischen den Werten in Spalten 5 und 9
6, 8, 10: K. ϱ_{ik} nach der Bayes'schen Statistik
Bemerkung b in Spalte 4: Wert in Spalte 8 liegt nicht zwischen den Werten in Spalten 6 und 10
fett: Werte der Spalten 5 bis 10 mit jeweils maximalem Betrag

1	2	3	4	5	6	7	8	9	10
lfd.	M-/U-	M-/U-	Bem.	— **JA** —		— **JC** —		— **JB** —	
Nr.	Nr. i	Nr. k	Val.	r_{ik}	ϱ_{ik}	r_{ik}	ϱ_{ik}	r_{ik}	ϱ_{ik}
1	1 / 1	2 / 2		0.306	0.296	0.218	0.211	0.054	0.052
2	1 / 1	3 / 3		0.140	0.134	0.247	0.240	0.286	0.278
3	1 / 1	4 / 4	ab	−0.051	−0.046	−0.042	−0.037	−0.100	−0.085
4	1 / 1	4 / 5		−0.053	−0.047	–	–	–	–
5	1 / 1	4 / 6		−0.117	−0.102	0.005	0.004	0.149	0.128
6	1 / 1	5 / 4		−0.127	−0.112	−0.029	−0.026	0.052	0.047
7	1 / 1	6 / 4		−0.192	−0.174	−0.027	−0.024	0.157	0.144
8	1 / 1	7 / 4		0.011	0.009	0.121	0.105	0.199	0.176
9	1 / 1	7 / 5		0.045	0.040	0.132	0.117	0.228	0.200
10	1 / 1	8 / 5		−0.023	−0.021	0.101	0.092	0.269	0.243
11	1 / 1	8 / 6		−0.134	−0.120	−0.019	−0.017	0.116	0.103
12	1 / 1	9 / 5		0.142	0.127	0.212	0.189	0.283	0.251
13	1 / 1	9 / 6		−0.017	−0.015	0.028	0.024	0.055	0.047
14	1 / 1	10 / 6		−0.160	−0.145	−0.056	−0.050	0.098	0.086
15	2 / 2	3 / 3		0.254	0.241	0.276	0.265	0.325	0.314
16	2 / 2	4 / 4		−0.291	−0.262	−0.165	−0.146	0.038	0.032
17	2 / 2	4 / 5		−0.091	−0.080	–	–	–	–
18	2 / 2	4 / 6		−0.140	−0.121	−0.056	−0.048	0.070	0.059
19	2 / 2	5 / 4		−0.165	−0.144	−0.056	−0.049	0.089	0.080
20	2 / 2	6 / 4		−0.054	−0.048	−0.002	−0.002	0.059	0.054
21	2 / 2	7 / 4		−0.239	−0.200	−0.059	−0.051	0.159	0.140
22	2 / 2	7 / 5		0.001	0.001	0.091	0.080	0.231	0.201
23	2 / 2	8 / 5		0.042	0.038	0.039	0.035	0.013	0.012
24	2 / 2	8 / 6		−0.135	−0.120	−0.069	−0.061	0.026	0.023
25	2 / 2	9 / 5	a	0.153	0.136	0.154	0.136	0.137	0.121
26	2 / 2	9 / 6		0.063	0.054	0.045	0.039	−0.006	−0.005
27	2 / 2	10 / 6		−0.152	−0.137	−0.115	−0.102	−0.054	−0.047
28	3 / 3	4 / 4		−0.088	−0.078	0.122	0.108	0.219	0.186
29	3 / 3	4 / 5		0.134	0.117	–	–	–	–

30	3 / 3	4 / 6		−0.014	−0.012	0.190	0.164	0.339	0.290
31	3 / 3	5 / 4		−0.025	−0.021	0.169	0.149	0.245	0.220
32	3 / 3	6 / 4		−0.000	−0.000	0.137	0.125	0.177	0.162
33	3 / 3	7 / 4		0.050	0.042	0.248	0.214	0.291	0.256
34	3 / 3	7 / 5		0.135	0.117	0.217	0.190	0.235	0.206
35	3 / 3	8 / 5	b	0.185	0.164	0.183	0.165	0.101	0.091
36	3 / 3	8 / 6		0.090	0.079	0.250	0.222	0.375	0.331
37	3 / 3	9 / 5		0.136	0.119	0.259	0.230	0.334	0.295
38	3 / 3	9 / 6		0.028	0.024	0.225	0.193	0.385	0.329
39	3 / 3	10 / 6		−0.001	−0.001	0.153	0.137	0.316	0.276
40	4 / 4	4 / 5	k	0.154	0.128	–	–	–	–
41	4 / 4	4 / 6	k	0.264	0.215	0.342	0.272	0.442	0.332
42	4 / 4	5 / 4	d	0.618	0.507	0.674	0.550	0.748	0.589
43	4 / 4	6 / 4	db	0.451	0.381	0.473	0.394	0.481	0.387
44	4 / 4	7 / 4	d	0.651	0.512	0.636	0.504	0.599	0.464
45	4 / 4	7 / 5		0.108	0.089	0.244	0.197	0.441	0.338
46	4 / 4	8 / 5		−0.111	−0.093	0.099	0.082	0.441	0.349
47	4 / 4	8 / 6		0.129	0.107	0.247	0.202	0.415	0.321
48	4 / 4	9 / 5		−0.020	−0.016	0.137	0.112	0.360	0.279
49	4 / 4	9 / 6		−0.001	−0.001	0.148	0.117	0.368	0.276
50	4 / 4	10 / 6		0.118	0.100	0.199	0.163	0.336	0.257
51	4 / 5	4 / 6	k	0.733	0.583	–	–	–	–
52	4 / 5	5 / 4		0.148	0.119	–	–	–	–
53	4 / 5	6 / 4		0.122	0.101	–	–	–	–
54	4 / 5	7 / 4		0.310	0.239	–	–	–	–
55	4 / 5	7 / 5	d	**0.865**	0.698	–	–	–	–
56	4 / 5	8 / 5	d	0.659	0.544	–	–	–	–
57	4 / 5	8 / 6		0.692	0.563	–	–	–	–
58	4 / 5	9 / 5	d	0.612	0.499	–	–	–	–
59	4 / 5	9 / 6		0.576	0.453	–	–	–	–
60	4 / 5	10 / 6		0.720	0.595	–	–	–	–
61	4 / 6	5 / 4		0.225	0.178	0.297	0.237	0.370	0.293
62	4 / 6	6 / 4		0.186	0.151	0.255	0.208	0.326	0.263
63	4 / 6	7 / 4	ab	0.302	0.229	0.316	0.244	0.299	0.232
64	4 / 6	7 / 5		0.635	0.505	0.607	0.479	0.545	0.420
65	4 / 6	8 / 5		0.435	0.354	0.475	0.384	0.521	0.415
66	4 / 6	8 / 6	d	0.780	0.625	0.805	0.640	0.837	0.651
67	4 / 6	9 / 5		0.350	0.280	0.416	0.331	0.498	0.389
68	4 / 6	9 / 6	d	0.600	0.465	0.623	0.479	0.645	0.487
69	4 / 6	10 / 6	d	0.816	0.664	0.814	0.651	0.813	0.627
70	5 / 4	6 / 4	d	0.850	0.696	0.827	0.691	0.790	0.671
71	5 / 4	7 / 4	db	0.698	0.533	0.696	0.553	0.674	0.551
72	5 / 4	7 / 5		0.132	0.106	0.304	0.246	0.522	0.423
73	5 / 4	8 / 5		−0.070	−0.057	0.194	0.161	0.572	0.478
74	5 / 4	8 / 6		0.104	0.084	0.210	0.171	0.328	0.268
75	5 / 4	9 / 5		0.064	0.052	0.236	0.193	0.449	0.368
76	5 / 4	9 / 6	ab	0.193	0.151	0.213	0.168	0.205	0.162
77	5 / 4	10 / 6		0.167	0.137	0.207	0.170	0.255	0.206
78	6 / 4	7 / 4	d	0.670	0.527	0.727	0.591	0.782	0.652
79	6 / 4	7 / 5		0.141	0.116	0.265	0.220	0.417	0.344
80	6 / 4	8 / 5		−0.072	−0.061	0.147	0.125	0.456	0.389
81	6 / 4	8 / 6		0.096	0.080	0.180	0.150	0.270	0.226
82	6 / 4	9 / 5		0.068	0.056	0.215	0.180	0.396	0.331
83	6 / 4	9 / 6		0.178	0.143	0.153	0.124	0.082	0.066

84	6 / 4	10 / 6	ab	0.168	0.142	0.174	0.146	0.166	0.137
85	7 / 4	7 / 5	kab	0.285	0.219	0.304	0.240	0.292	0.232
86	7 / 4	8 / 5		−0.013	−0.010	0.130	0.105	0.279	0.229
87	7 / 4	8 / 6		0.166	0.129	0.208	0.165	0.221	0.177
88	7 / 4	9 / 5		0.154	0.119	0.226	0.180	0.275	0.221
89	7 / 4	9 / 6		0.094	0.071	0.138	0.106	0.148	0.115
90	7 / 4	10 / 6		0.207	0.163	0.192	0.153	0.150	0.119
91	7 / 5	8 / 5	d	0.739	0.610	0.767	0.631	0.806	0.656
92	7 / 5	8 / 6		0.648	0.527	0.604	0.488	0.514	0.409
93	7 / 5	9 / 5	d	0.687	0.559	0.727	0.589	0.782	0.623
94	7 / 5	9 / 6		0.567	0.446	0.520	0.408	0.421	0.325
95	7 / 5	10 / 6		0.652	0.538	0.606	0.493	0.515	0.406
96	8 / 5	8 / 6	k	0.650	0.541	0.627	0.521	0.573	0.470
97	8 / 5	9 / 5	d	0.663	0.552	0.713	0.593	0.789	0.649
98	8 / 5	9 / 6		0.458	0.369	0.447	0.360	0.402	0.320
99	8 / 5	10 / 6	b	0.496	0.419	0.504	0.421	0.510	0.415
100	8 / 6	9 / 5		0.430	0.353	0.461	0.376	0.488	0.393
101	8 / 6	9 / 6	d	0.564	0.448	0.610	0.482	0.672	0.523
102	8 / 6	10 / 6	d	0.850	**0.708**	**0.866**	**0.711**	**0.899**	**0.715**
103	9 / 5	9 / 6	k	0.520	0.413	0.515	0.407	0.484	0.378
104	9 / 5	10 / 6		0.352	0.293	0.388	0.319	0.443	0.353
105	9 / 6	10 / 6	d	0.644	0.519	0.662	0.526	0.691	0.533

Tabelle A.9: Korrelationskoeffizienten zur Konstruktvalidität der AC-Serie ST

Legende siehe Tabelle A.8, Spalten 2 und 3 nach Tabelle 5.3

1 lfd. Nr.	2 M-/U-Nr. i	3 M-/U-Nr. k	4 Val.	5 — ST — r_{ik}	6 — ST — ϱ_{ik}
1	1 / 1	1 / 5	k	0.460	0.157
2	1 / 1	2 / 1	d	0.538	0.255
3	1 / 1	3 / 3		0.317	0.183
4	1 / 1	3 / 4		0.335	0.300
5	1 / 1	4 / 3		0.487	0.302
6	1 / 1	5 / 3		−0.055	−0.030
7	1 / 1	6 / 2		0.290	0.132
8	1 / 1	6 / 6		0.096	0.054
9	1 / 1	6 / 7		0.053	0.033
10	1 / 1	7 / 2		0.226	0.150
11	1 / 1	7 / 6		0.263	0.174
12	1 / 1	7 / 7		0.243	0.123
13	1 / 1	8 / 2		0.384	0.289
14	1 / 1	8 / 6		0.062	0.038
15	1 / 1	8 / 7		0.141	0.107
16	1 / 5	2 / 1		−0.020	−0.004
17	1 / 5	3 / 3		0.117	0.028
18	1 / 5	3 / 4		0.448	0.167
19	1 / 5	4 / 3		0.146	0.038
20	1 / 5	5 / 3		0.129	0.029
21	1 / 5	6 / 2		0.193	0.037
22	1 / 5	6 / 6		0.018	0.004
23	1 / 5	6 / 7		0.168	0.044
24	1 / 5	7 / 2		0.235	0.065
25	1 / 5	7 / 6		0.331	0.091
26	1 / 5	7 / 7		0.144	0.030
27	1 / 5	8 / 2		0.367	0.115
28	1 / 5	8 / 6		−0.028	−0.007
29	1 / 5	8 / 7		0.157	0.049
30	2 / 1	3 / 3		0.248	0.083
31	2 / 1	3 / 4		0.009	0.005
32	2 / 1	4 / 3		0.551	0.198
33	2 / 1	5 / 3		0.138	0.044
34	2 / 1	6 / 2		0.136	0.036
35	2 / 1	6 / 6		0.065	0.021
36	2 / 1	6 / 7		0.176	0.064
37	2 / 1	7 / 2		0.297	0.114
38	2 / 1	7 / 6		0.348	0.133
39	2 / 1	7 / 7		0.214	0.063
40	2 / 1	8 / 2		0.422	0.184
41	2 / 1	8 / 6		0.140	0.050
42	2 / 1	8 / 7		0.311	0.136
43	3 / 3	3 / 4	k	0.660	0.417
44	3 / 3	4 / 3	d	0.592	0.259
45	3 / 3	5 / 3	d	0.109	0.042
46	3 / 3	6 / 2		0.377	0.121

47	3 / 3	6 / 6		0.415	0.165
48	3 / 3	6 / 7		0.485	0.214
49	3 / 3	7 / 2		−0.017	−0.008
50	3 / 3	7 / 6		0.133	0.062
51	3 / 3	7 / 7		0.140	0.050
52	3 / 3	8 / 2		−0.045	−0.024
53	3 / 3	8 / 6		0.225	0.097
54	3 / 3	8 / 7		0.337	0.180
55	3 / 4	4 / 3		0.458	0.311
56	3 / 4	5 / 3		0.147	0.088
57	3 / 4	6 / 2		0.344	0.172
58	3 / 4	6 / 6		0.267	0.165
59	3 / 4	6 / 7		0.254	0.174
60	3 / 4	7 / 2		−0.051	−0.037
61	3 / 4	7 / 6		0.045	0.033
62	3 / 4	7 / 7		0.369	0.204
63	3 / 4	8 / 2		0.007	0.005
64	3 / 4	8 / 6		0.061	0.041
65	3 / 4	8 / 7		0.123	0.102
66	4 / 3	5 / 3	d	0.035	0.015
67	4 / 3	6 / 2		0.309	0.107
68	4 / 3	6 / 6		0.515	0.220
69	4 / 3	6 / 7		0.211	0.100
70	4 / 3	7 / 2		0.536	0.270
71	4 / 3	7 / 6		0.100	0.050
72	4 / 3	7 / 7		0.321	0.123
73	4 / 3	8 / 2		0.574	0.328
74	4 / 3	8 / 6		0.368	0.171
75	4 / 3	8 / 7		0.310	0.178
76	5 / 3	6 / 2		0.077	0.023
77	5 / 3	6 / 6		−0.026	−0.010
78	5 / 3	6 / 7		−0.028	−0.012
79	5 / 3	7 / 2		−0.090	−0.040
80	5 / 3	7 / 6		−0.018	−0.008
81	5 / 3	7 / 7		0.233	0.079
82	5 / 3	8 / 2		−0.046	−0.023
83	5 / 3	8 / 6		−0.126	−0.052
84	5 / 3	8 / 7		0.057	0.029
85	6 / 2	6 / 6	k	0.239	0.075
86	6 / 2	6 / 7	k	0.194	0.068
87	6 / 2	7 / 2	d	−0.041	−0.015
88	6 / 2	7 / 6		0.388	0.143
89	6 / 2	7 / 7		0.220	0.062
90	6 / 2	8 / 2	d	0.162	0.068
91	6 / 2	8 / 6		0.214	0.073
92	6 / 2	8 / 7		0.155	0.065
93	6 / 6	6 / 7	k	0.346	0.149
94	6 / 6	7 / 2		0.467	0.213
95	6 / 6	7 / 6	d	−0.006	−0.003
96	6 / 6	7 / 7		0.540	0.188
97	6 / 6	8 / 2		0.507	0.263
98	6 / 6	8 / 6	d	0.798	0.337
99	6 / 6	8 / 7		0.579	0.302
100	6 / 7	7 / 2		0.112	0.057

101	6 / 7	7 / 6		0.425	0.215
102	6 / 7	7 / 7	d	0.214	0.083
103	6 / 7	8 / 2		0.135	0.078
104	6 / 7	8 / 6		0.350	0.164
105	6 / 7	8 / 7	d	0.811	0.469
106	7 / 2	7 / 6	k	−0.052	−0.028
107	7 / 2	7 / 7	k	0.294	0.120
108	7 / 2	8 / 2	d	**0.891**	**0.544**
109	7 / 2	8 / 6		0.452	0.224
110	7 / 2	8 / 7		0.449	0.275
111	7 / 6	7 / 7	k	−0.141	−0.058
112	7 / 6	8 / 2		0.173	0.105
113	7 / 6	8 / 6	d	0.285	0.141
114	7 / 6	8 / 7		0.304	0.186
115	7 / 7	8 / 2		0.412	0.191
116	7 / 7	8 / 6		0.261	0.099
117	7 / 7	8 / 7	d	0.467	0.218
118	8 / 2	8 / 6	k	0.458	0.258
119	8 / 2	8 / 7	k	0.432	0.301
120	8 / 6	8 / 7	k	0.586	0.332

Tabelle A.10: Teilnehmerbezogene Korrelationskoeffizienten der AC-Serien JA, JB, JC und ST

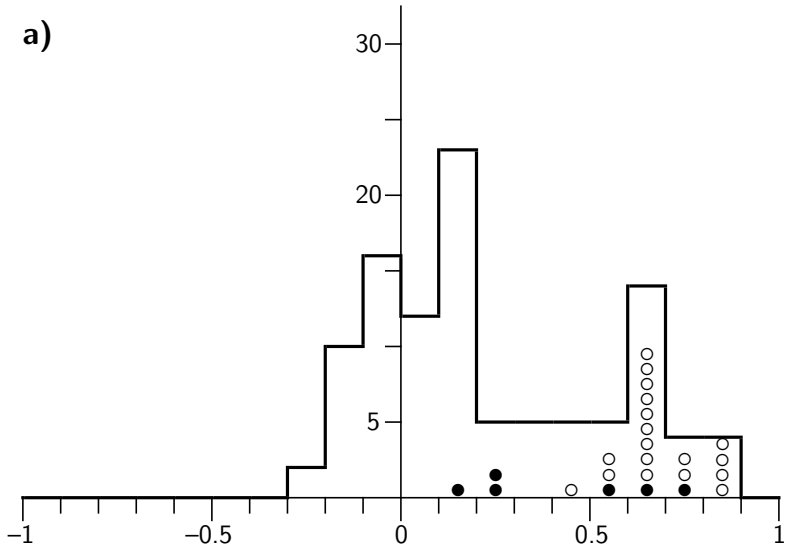
Die Tabellenwerte stehen auf der folgenden Seite.

Bedeutung der Spalten:
1: laufende Zeilen-Nummer
2 und 3: Merkmals-Nummer (M-Nr.) des Partners j bzw. des Partners l des Merkmal-Paares (j, l) (siehe Tabellen 5.2 und 5.3)
4 bis 7: teilnehmerbezogene Korrelationskoeffizienten δ_{jl} zum Merkmal-Paar (j, l). Wenn exakt gleich 0, dann Angabe 0 ohne Dezimalstellen
fett: Werte der Spalten 4 bis 7 mit jeweils maximalem Betrag

1 lfd. Nr.	2 M-Nr. j	3 M-Nr. l	4 **JA** δ_{jl}	5 **JC** δ_{jl}	6 **JB** δ_{jl}	7 **ST** δ_{jl}
1	1	2	0	0	0	0
2	1	3	0	0	0	– 0.0329
3	1	4	0	0	0	0
4	1	5	0	0	0	0
5	1	6	0	0	0	0.0076
6	1	7	0	0	0	– 0.0050
7	1	8	0	0	0	0.0039
8	1	9	0	0	0	–
9	1	10	0	0	0	–
10	2	3	0	0	0	0
11	2	4	0	0	0	0
12	2	5	0	0	0	0
13	2	6	0	0	0	0
14	2	7	0	0	0	0
15	2	8	0	0	0	0
16	2	9	0	0	0	–
17	2	10	0	0	0	–
18	3	4	0	0	0	0
19	3	5	0	0	0	0
20	3	6	0	0	0	– 0.0100
21	3	7	0	0	0	0.0039
22	3	8	0	0	0	– 0.0127
23	3	9	0	0	0	–
24	3	10	0	0	0	–
25	4	5	0.0128	0.0203	0.0216	0
26	4	6	0.0043	0.0101	0.0158	0
27	4	7	**0.0875**	**0.0650**	0.0328	0
28	4	8	0.0026	0.0044	0.0109	0
29	4	9	0.0004	0.0034	0.0061	–
30	4	10	0.0198	0.0116	0.0144	–
31	5	6	0.0184	0.0351	**0.0620**	0
32	5	7	0.0158	0.0059	– 0.0096	0
33	5	8	0	0.0015	0.0039	0
34	5	9	0	0.0017	0.0045	–
35	5	10	0	– 0.0034	– 0.0086	–
36	6	7	0.0074	0.0118	0.0185	– 0.0163
37	6	8	0	– 0.0013	– 0.0034	0.0256
38	6	9	0	– 0.0005	– 0.0012	–
39	6	10	0	0.0033	0.0083	–
40	7	8	0.0044	0.0147	0.0301	**0.0405**
41	7	9	0.0067	0.0168	0.0319	–
42	7	10	0.0029	0.0085	0.0166	–
43	8	9	0.0171	0.0335	0.0588	–
44	8	10	0.0270	0.0204	0.0106	–
45	9	10	0.0005	– 0.0088	– 0.0227	–
	Mittelwerte $\overline{\delta}$:		0.0051	0.0063	0.0087	0.0057
	empir. Standardabw. zu $\overline{\delta}$:		0.0021	0.0018	0.0022	0.0020

Bild A.1: Histogramme der Korrelationskoeffizienten der AC-Serie JA

a) Korrelationskoeffizienten (K.) zur Konstruktvalidität
nach der konventionellen Statistik
b) K. zur Konstruktvalidität nach der Bayes'schen Statistik
•: K. zur konvergenten Konstruktvalidität
○: K. zur diskriminanten Konstruktvalidität



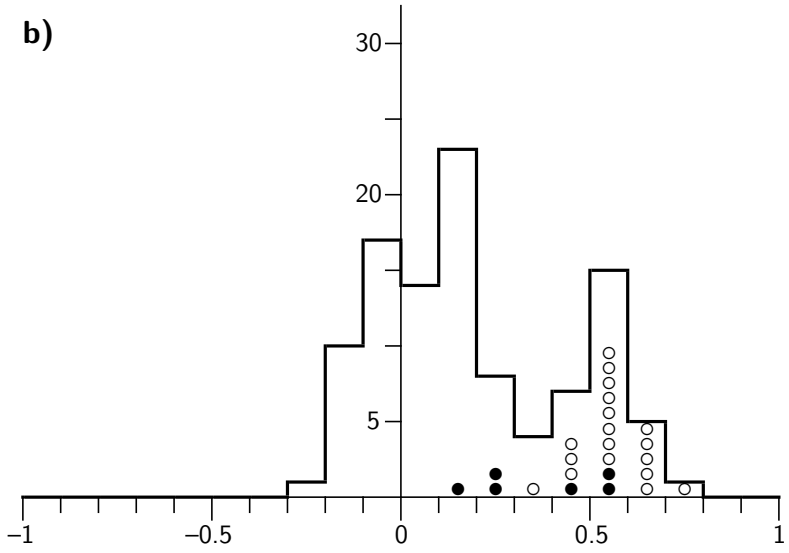


Bild A.2: Histogramme der Korrelationskoeffizienten der AC-Serie JB

a) Korrelationskoeffizienten (K.) zur Konstruktvalidität
nach der konventionellen Statistik
b) K. zur Konstruktvalidität nach der Bayes'schen Statistik
•: K. zur konvergenten Konstruktvalidität
○: K. zur diskriminanten Konstruktvalidität

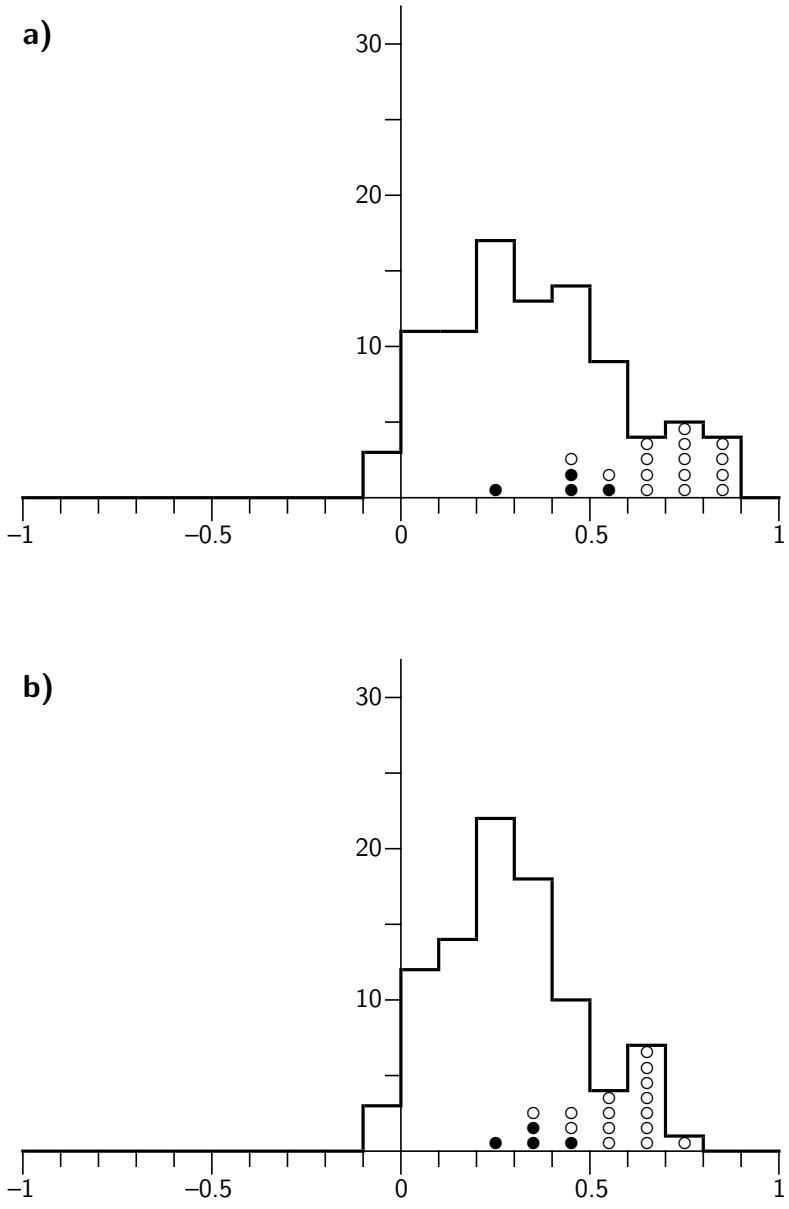


Bild A.3: Histogramme der Korrelationskoeffizienten der AC-Serie JC

a) Korrelationskoeffizienten (K.) zur Konstruktvalidität
nach der konventionellen Statistik
b) K. zur Konstruktvalidität nach der Bayes'schen Statistik
•: K. zur konvergenten Konstruktvalidität
○: K. zur diskriminanten Konstruktvalidität

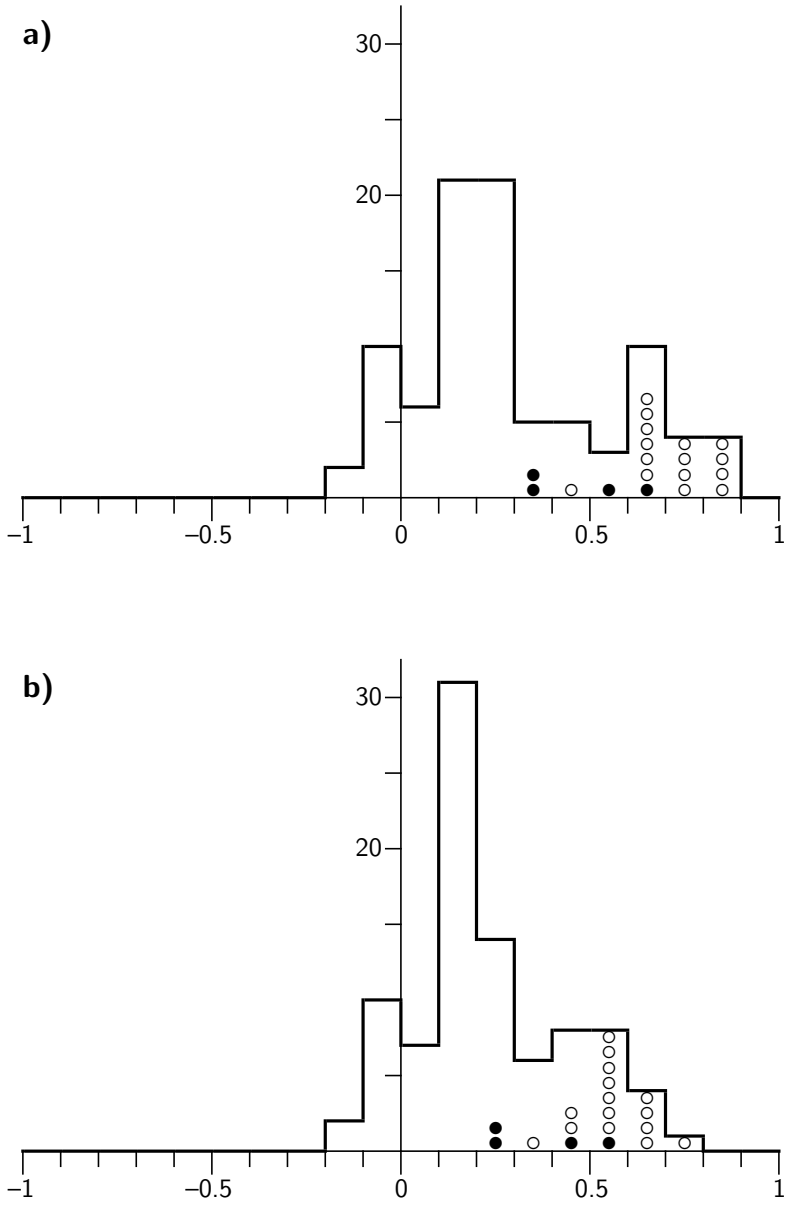
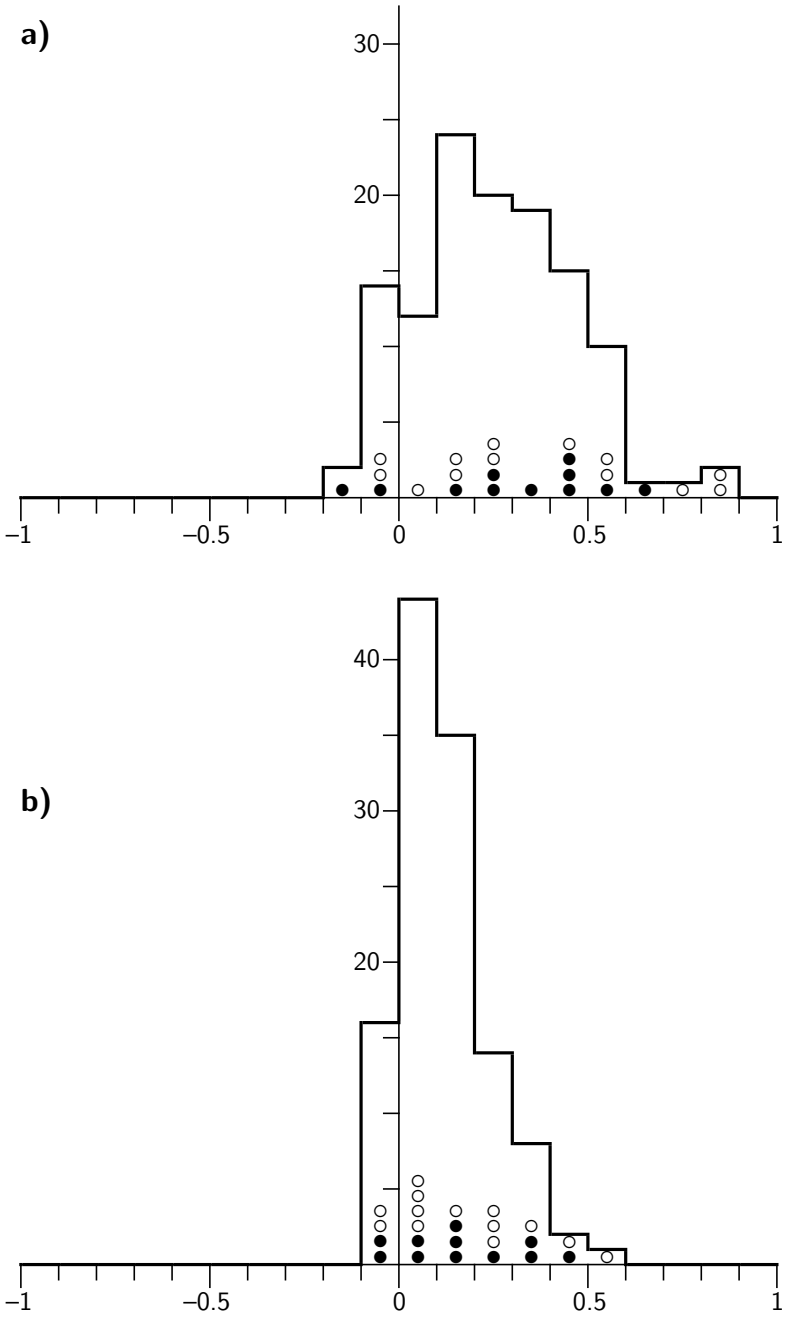


Bild A.4: Histogramme der Korrelationskoeffizienten der AC-Serie ST

a) Korrelationskoeffizienten (K.) zur Konstruktvalidität
nach der konventionellen Statistik
b) K. zur Konstruktvalidität nach der Bayes'schen Statistik
•: K. zur konvergenten Konstruktvalidität
○: K. zur diskriminanten Konstruktvalidität



Anhang B: Beschreibung des Auswahlprogramms QWAHL

Diese Beschreibung des Auswahlprogramms QWAHL ist so verfasst, dass sie auch unabhängig vom Hauptteil dieser Arbeit für die Handhabung des Programms verwendet werden kann. Das Programm kann nicht nur im Personalbereich, sondern auch ganz allgemein bei beliebigen AC-analogen Urteils- und Entscheidungsprozessen eingesetzt werden. Deshalb wird in der Programmbeschreibung abweichend vom Hauptteil dieser Arbeit anstelle des Wortes „Teilnehmer“ das Wort „Alternative“ benutzt. Das Programm QWAHL wurde auf der Grundlage des in dieser Arbeit vorgeschlagenen Auswerteverfahrens in Zusammenarbeit mit einer auf dem Gebiet der Personalentwicklung tätigen Unternehmensberatung erstellt. Derzeit liegt das Programm erst in der beschriebenen Experimentier- und Demonstrationsversion vor. Diese bietet zwar alle wesentlichen Funktionen und kann durchaus in der Praxis angewendet werden, ist aber nicht völlig ausgereift, insbesondere ist die Handhabung für den Routineeinsatz noch nicht bequem genug und es fehlen noch einige im Hauptteil erwähnte sinnvolle Eigenschaften.

Die Daten des Anwendungsbeispiels (Anlage 1) zum Programm QWAHL stammen im Wesentlichen aus einem realen AC. Sie wurden lediglich anonymisiert und nur wenig ergänzt und geändert, um einen geeigneten Datensatz für die Entwicklung, den Test und die Demonstration des Programms zu bilden.

Beschreibung des Auswahlprogramms QWAHL

Gültig für Programmversion 3.09.

Inhaltsverzeichnis

1 Einführung und Übersicht

1.1 Allgemeine Charakterisierung des Programms

Das Programm QWAHL unterstützt ganz allgemein die Bewertung, den Vergleich und die Auswahl von Alternativen und damit auch die Entscheidungsfindung durch systematische und konsequente Analyse vorliegender, auch unvollständiger Information anhand von Vorgaben.

Die begleitende wissenschaftliche Untersuchung bildet die theoretische, Bayes-statistische Grundlage für das im Programm angewendete Auswahlverfahren.

Die einzelnen Bewertungen der in Betracht gezogenen Alternativen (z.B. Teilnehmer eines Assessment Centers (AC), Kandidaten in einer Prüfung, gleichartige materielle oder ideelle Objekte wie Gebäude bzw. Vorgehensweisen) hinsichtlich der für die Auswahl wichtigen Eigenschaften durch Beobachter (z.B. Gutachter, Prüfer, Beisitzer, Experimentatoren) unter verschiedenen Übungen (Tests, d.h. Untersuchungen unter vorgegebenen Bedingungen, Aufgaben und Verfahren; z.B. Versuche, Prüfungen, Experimente) werden anhand gewichteter Merkmale ausgewertet und zu Gesamtbewertungen der einzelnen Alternativen zusammengefasst. Auch Teilauswertungen zu wichtigen, die Alternativen betreffenden Aspekten sowie eine Evaluierung der Übungen und Beobachter sind möglich. Die Ergebnisse werden nummerisch und graphisch dargestellt und können so leicht miteinander verglichen werden. Das Besondere des Programms ist, dass neben den Bewertungen der Alternativen selbst auch ihre auf Mangel an Information beruhende Unsicherheit auf der Grundlage der Bayes'schen Statistik erfasst bzw. berechnet wird. Dadurch können die Qualität der Ergebnisse und das Vertrauen darin begründet werden und es ist möglich, beim kritischen Vergleich der Alternativen auch das Risiko der Auswahl einer Alternative quantitativ einzuschätzen (Abschnitt 3.7).

Der Name QWAHL des Auswahlprogramms ist entstanden durch lautliches und schriftliches Verschmelzen der Wörter „Wahl“ und „Qual“ aus dem Sprichwort „Wer die Wahl hat, hat die Qual“.

1.2 Kurzbeschreibung des Auswahlverfahrens

Jede Aufgabe, Alternativen zu bewerten, miteinander zu vergleichen und sich gegebenenfalls für eine davon zu entscheiden, wird anhand einer Reihe von Merkmalen gelöst. Diese Merkmale entsprechen genau zu definierenden Eigenschaften der Alternativen.

Die Ausprägungen der Merkmale bei den einzelnen Alternativen werden aufgrund der in einigen Übungen gesammelten Information von Beobachtern auf einer vereinbarten Skala jeweils einzeln bewertet. Da die Übungen unterschiedlich aussagekräftig für die einzelnen Eigenschaften sind und auch die Merkmale unterschiedlich wichtig für die Auswahl sein können, sind ihnen Gewichte zuzuordnen. Die Einzelbewertungen der Beobachter werden so gewichtet zusammengeführt und ergeben Bewertungen zu den einzelnen Merkmalen, gegebenenfalls auch zu Gruppen von Merkmalen, die wichtigen Teilaspekten des Vergleichs und der Auswahl entsprechen. Schließlich erfolgen die Gesamtbewertungen der Alternativen. Ergebnisse einer Vorauswahl können ebenfalls gewichtet Berücksichtigung finden.

Die Einzelbewertungen der Beobachter in den Übungen bezüglich der Ausprägung der Merkmale der Alternativen sind mehr oder weniger unsicher, weil sehr oft nicht genügend Information gesammelt werden kann. Deshalb dürfen die Beobachter, dieser Tatsache entsprechend, zu einer Eigenschaft jeweils eine minimale und eine maximale Bewertung angeben, die sie nach vernünftiger und realistischer Einschätzung der vorliegenden Information noch als möglicherweise zutreffend erachten. Diese Angabe der minimalen und der maximalen Bewertungen, die natürlich auch übereinstimmen dürfen, gestattet es schließlich, zu jeder errechneten Bewertung jeweils auch die zugehörige Unsicherheit zu quantifizieren (Abschnitt 2.4). Auch eine gegebene Unsicherheit der Gewichte kann global berücksichtigt werden. Der Begriff und die Quantifizierung der Unsicherheit basieren auf der Bayes'schen Statistik und stammen aus der physikalischen Messtechnik (Abschnitt 3.7).

Weil die Information über die Merkmale der Alternativen in Form minimaler und maximaler Einzelbewertungen durch die Beobachter vorliegt, kann im Rahmen der Bayes'schen Statistik die Korrelation zwischen den Merkmalen einer Alternative bei der Berechnung der Unsicherheiten vernachlässigt werden. Dies mag verwundern, wird aber in der begleitenden Untersuchung ausführlich begründet.

1.3 Hinweise zur Programmversion

Die vorliegende Beschreibung des Programms QWAHL bezieht sich auf die Version 3.09. Diese ist noch eine reine Experimentier- und Demonstrationsversion. Sie bietet zwar alle wesentlichen Funktionen und kann durchaus in der Praxis angewendet werden, ist aber nicht völlig ausgereift, insbesondere ist ihre Handhabung noch nicht bequem genug. Obwohl so sorgfältig wie möglich erstellt und getestet, kann sie noch Fehler enthalten. Eine Gewähr auf Fehlerfreiheit kann deshalb nicht übernommen werden. Anregungen und Kritik sind willkommen.

Der Quellcode des Programms QWAHL wurde in der Programmiersprache BASIC im Dialekt (Programmierumgebung) von QBASIC (Quick BASIC) unter dem Betriebssystem DOS entwickelt und liegt als ASCII-Datei QWAHL-03.BAS für DOS vor. Bearbeitung und Ablauf in der Programmierumgebung von Visual BASIC unter DOS oder Windows sind ebenfalls möglich. Elemente der objektorientierten Programmierung zur Gestaltung der Bedienung und Graphik wurden noch nicht verwendet ♠ [1].

Zum Programm QWAHL gehört der ablauffähige Code in der Datei QWAHL-03.EXE, der BASIC-Quellcode in der Datei QWAHL-03.BAS und eine Beispiel-Eingabedatei QWAHL.EIN. Die beiden letzteren sind ASCII-Dateien im DOS-Format. Die Beispiel-Eingabedatei beschreibt eine konkrete, in der Unternehmenspraxis häufige Entscheidungsaufgabe, aus mehreren Stellenbewerbern, die einem Assessment Center (AC) unterzogen werden, den hinsichtlich gewichteter Merkmale bestgeeigneten Bewerber auszuwählen. Die Daten in dieser Datei stammen im Wesentlichen aus einem realen AC. Sie wurden nur wenig ergänzt und geändert, um einen für Programmentwicklung und -test geeigneten Beispiel-Datensatz zu bilden.

1.4 Systemvoraussetzungen und Programmstart

Das Programm QWAHL läuft auf den verbreiteten IBM-kompatiblen Personal Computern (PC) mit VGA-Bildschirm und Standarddrucker unter den Betriebssystemen DOS 6.22 oder Windows 3.11, 95, 98, Me ohne besondere weitere Systemvorausetzungen. Andere DOS- oder Windows-Versionen dürften keine Schwierigkeiten bereiten. Es läuft auch auf Apple-Macintosh-Computern unter dem Betriebssystem Mac OS 9 in der PC-Simulation Virtual PC.

Der Programmstart erfolgt unter DOS oder unter Windows im DOS-Fenster (MS-DOS-Eingabeaufforderung) durch Eingabe des Programmnamens QWAHL-03 am DOS-Prompt. Vorher sind das Laufwerk und das Verzeichnis (Ordner) einzustellen, in denen sich der ablauffähige Code QWAHL-03.EXE befindet. Liegt dieser z.B. auf dem Laufwerk A im Verzeichnis AUSWAHL, so sind am DOS-Prompt die Befehle A: und CD AUSWAHL einzugeben. Unter Windows kann das Programm auch mit einer Maus durch Doppelklicken auf die Datei QWAHL-03.EXE im Dateimanager gestartet werden. Das Programm kann auch als Quellcode QWAHL-03.BAS in die Programmierumgebungen QBASIC oder Visual BASIC geladen, dort betrachtet, nötigenfalls geändert und durch Drücken der Tastenkombination Shift-F5 gestartet werden.

[1] Das Zeichen ♠ weist auf später vorzusehende Programmerweiterungen hin.

Das Programm benötigt zum Ablauf immer eine ASCII-Eingabedatei mit dem Namen QWAHL.EIN in demselben Verzeichnis. Diese Datei definiert die aktuelle Auswahlaufgabe und enthält neben allen Parameterdaten auch die Einzelbewertungen der Beobachter. Sie kann mit einem beliebigen geeigneten Editorprogramm erstellt, betrachtet oder geändert werden. Wie dies zu geschehen hat, wird in Abschnitt 2 beschrieben.

1.5 Programmfunktionen und Ergebnisausgabe

Die Ausgabe der vom Programm QWAHL errechneten Ergebnisse zu den einzelnen Alternativen erfolgt mit den jeweils zugehörigen Unsicherheiten und Rängen in nummerischer und graphischer Darstellung auf den Bildschirm oder in zwar vereinfachter, aber vollständiger nummerischer Darstellung auch in eine Datei oder mittels eines Druckers auf Papier. Die Ausgabe auf den Bildschirm ist in Einzelbildern und Bildfolgen organisiert, die den unterschiedlichen Programmfunktionen und Fragestellungen entsprechen und zwischen denen direkt oder über ein Menü leicht hin und her geschaltet werden kann (Abschnitt 3.2).

Neben einem Menübild, das auch eine Bedienungshilfe zeigt, gibt es die folgenden Funktionen, d.h. Bilder und Bildfolgen, die in Abschnitt 3 genauer beschrieben werden:

- Nummerischer und graphischer Vergleich der Alternativen hinsichtlich ihrer 1) Gesamtergebnisse, 2) Vorauswahlergebnisse, wenn vorhanden, sowie 3) Endergebnisse, d.h. der zusammengefassten Gesamtergebnisse und Vorauswahlergebnisse, wenn letztere vorliegen.
- 4) Nummerischer Vergleich der Alternativen hinsichtlich aller ihrer Ränge bei den Gesamt-, Vorauswahl- und Endergebnissen sowie bei den Merkmalsgruppen, Merkmalen und Übungen.
- Nummerischer und graphischer Vergleich der Alternativen hinsichtlich 5) jeder einzelnen Merkmalsgruppe, 6) jedes einzelnen Merkmals und 7) jeder einzelnen Übung.
- 8) Nummerische und 9) graphische Darstellung aller Ergebnisse jeder einzelnen Alternative.

Die Funktionen sind hier nach ihren Funktionsnummern aufgezählt. Eine Funktion wird aufgerufen durch Drücken der Zifferntaste, die der Funktionsnummer entspricht. Das Bild der Funktion 1 (Gesamtergebnisse) erscheint sofort nach Beendigung aller Berechnungen.

Man beachte den Unterschied zwischen Gesamtergebnissen (Funktion 1) und Endergebnissen (Funktion 3). In die Gesamtergebnisse sind die Vorauswahlergebnisse noch

n i c h t eingeflossen. Gesamtergebnisse und Endergebnisse sind identisch, wenn Vorauswahlergebnisse nicht vorliegen. Die Vorauswahlergebnisse sind die Endergebnisse einer vorausgegangenen anderen Anwendung des Programms QWAHL mit denselben Alternativen und derselben Bewertungsskala (Abschnitt 2.3, Punkt 3), aber mit im Allgemeinen (jedoch nicht unbedingt) anderen Merkmalen, Übungen und Beobachtern.

Die Datei- und Druckausgabe erfolgt in einer Form ähnlich der Bildfolge der Funktion 8 (nummerische Darstellung aller Ergebnisse jeder Alternative, Anlage 2). Ist ein geeigneter Farbdrucker vorhanden, könnte auch die Ausgabe der graphischen Bilder direkt vom Bildschirm gelingen, unter DOS durch Drücken der Druck-Taste. Hierbei muss aber mit Schwierigkeiten gerechnet werden. Unter Windows kann ein Bild als Bildschirmphoto (screen shot) festgehalten und mittels eines Graphikbetrachtungsprogramms bearbeitet und gedruckt werden.

Auch die Aussagefähigkeit der Übungen bezüglich der einzelnen Merkmale kann überprüft werden (Abschnitt 3.7), ebenso lässt sich die Urteilsfähigkeit der Beobachter vergleichen. Diese Funktionen sind jedoch noch nicht implementiert ♠.

2 Dateneingabe

2.1 Allgemeines

Das Programm QWAHL benötigt zum Ablauf immer eine ASCII-Eingabedatei mit dem Namen QWAHL.EIN in demselben Verzeichnis. Diese Datei definiert die aktuelle Auswahlaufgabe und enthält neben den Daten zu allen Parametern (Variablen) auch die Einzelbewertungen der Beobachter. Sie kann mit einem beliebigen geeigneten Editorprogramm erstellt, geöffnet, betrachtet oder geändert werden, z.B. mit dem DOS-Editor, der sich durch Eingeben des Befehls EDIT QWAHL.EIN am DOS-Prompt aufrufen lässt, oder mit WORD unter Windows als MS-DOS-Text. Um die folgende Beschreibung der Dateneingabe besser zu verstehen, betrachte man die beigefügte Beispiel-Eingabedatei QWAHL.EIN (Abschnitt 1.4 und Anlage 1). Die Eingabedatei sollte zunächst unter einem beliebig gewählten Namen erstellt und für den Programmablauf als Kopie davon unter dem Namen QWAHL.EIN bereitgestellt werden. Das ist zweckmäßig, weil nach Ablauf des Programms oft einige Parameterdaten, z.B. Gewichte von Merkmalen, zu ändern sind, um nach abermaligem Ablauf den Einfluss dieser Änderung auf das Ergebnis festzustellen. Solche Änderungen sollten nur in der Eingabedatei QWAHL.EIN vorgenommen werden. Die ursprüngliche Eingabedatei bleibt dadurch dann unberührt.

2.2 Aufbau der Eingabedatei

Die Eingabedatei besteht aus einer Folge von Blöcken, jeder dieser Blöcke aus einem Kommentarteil und einem Datenteil. Auf den letzten Block darf beliebiger Kommentar folgen. Die Datei darf kein Tabulatorzeichen enthalten (Abschnitt 2.5).

Nur die letzte Zeile jedes Kommentarteils darf und muss mit dem Zeichen $*$ beginnen, woran das Programm das Kommentarende erkennt. Ein Kommentarteil darf ansonsten beliebig gestaltet werden, er wird vom Programm überlesen. Er sollte den folgenden Datenteil erläutern, der erste Kommentarteil auch die ganze aktuelle Auswahlaufgabe.

Ein Datenteil besteht aus Datenzeilen mit einer festgelegten Folge von Daten und Zeichenfolgen (Strings). In jeder Zeile sind die Daten durch Zwischenräume oder Komma, die Strings nur durch Komma voneinander zu trennen. Deshalb darf ein String selbst kein Komma enthalten. Bei den Daten ist als Dezimalzeichen ein Punkt zu verwenden, z.B. 3.2 statt 3,2. Die Daten sind als ganze Zahlen, Festpunkt- oder Gleitpunktzahlen anzugeben. Die Anzahl der zwischen den Datenzeilen eines Datenteils vorgesehenen Leerzeilen darf nicht verändert werden (wie im Beispiel in Anlage 1 bei den Bewertungen im Datenteil des zehnten Blocks, siehe auch Abschnitt 2.4).

2.3 Angabe der Auswahlparameter

Die Datenteile der ersten neun Blöcke der Eingabedatei enthalten die Daten der Parameter, die die Auswahlaufgabe genau beschreiben. Als Beispiel dient Anlage 1.

1) Der Datenteil des ersten Blocks besitzt nur eine Datenzeile mit einem einzigen String, dem Titel der Auswahlaufgabe. Dieser Titel sollte kurz, genau und prägnant sein. Im Beispiel wurde der Ortsname der Sparkasse, bei der das AC stattfand, geändert.

2) Der Datenteil des zweiten Blocks besitzt ebenfalls nur eine Datenzeile. Diese enthält die Daten zu den folgenden fünf Variablen (Anzahlen): Anzahl der Alternativen (4), Beobachter (3), Übungen (3), Merkmalsgruppen (3) und Merkmale (11) in dieser Reihenfolge. (In Klammern stehen die Werte des Beispiels.) Praxisgerechter wäre es, jeder Übung eine eigene Beobachteranzahl zuzuordnen. Dies ist jedoch noch nicht implementiert ♠.

Obwohl für die Berechnungen keine Einschränkungen erforderlich sind, sollten die Anzahlen aus Gründen des Aufwands und der Darstellung der Ergebnisse auf dem Bildschirm nicht größer als etwa 10 gewählt werden. Die Anzahlsumme der Merkmalsgruppen, Merkmale und Übungen sollte 19 nicht überschreiten, weil sonst die Bilder der

Funktionen 4, 8 und 9 nicht mehr auf den Bildschirm passen ♠. Zweckmäßig ist es, nur solche Alternativen zuzulassen, die alle unabdingbaren Voraussetzungen auch wirklich erfüllen oder in die engere Wahl gekommen sind, d.h. es sollte eine Vorauswahl getroffen werden. Wenn eine Entscheidung zu fällen ist, können Merkmale, die von allen Alternativen gleichermaßen erfüllt werden, unbeachtet bleiben, weil sie zur Entscheidungsfindung nichts beitragen können. Mehrere Merkmale von geringem Gewicht können mitunter zu einem umfassenden Merkmal zusammengelegt werden.

3) Auch der Datenteil des dritten Blocks besitzt nur eine Datenzeile. Diese enthält die Daten der folgenden vier Variablen zur Beschreibung der Bewertungsskala: die untere Grenze Su und die obere Grenze So der Skala, die Schrittweite dS der Skalenbeschriftung und einen Teiler eS dieser Schrittweite für die Skalenteilung. Obwohl für die Berechnungen keine Einschränkungen erforderlich sind, müssen die Skalendaten aus darstellungstechnischen Gründen für die Graphik ganzzahlig sein und die folgenden Bedingungen erfüllen:

- $-100 < \mathrm{Su} < \mathrm{So} < 1000$, d.h. Su und So sind mit Vorzeichen höchstens dreiziffrig zu wählen,
- $\mathrm{dS} > 0$ und $\mathrm{So} - \mathrm{Su} = \mathrm{dS} \cdot \mathrm{t}$. Es ist dS also ein Teiler der Skalenlänge So − Su. Für t sind nur die Werte 1 bis 8, 10 und 12 zulässig.
- $\mathrm{eS} > 0$. Es ist eS ein Teiler von dS. Zweckmäßig, aber nicht notwendig ist die Wahl $\mathrm{eS} \leq 5$ und $10 \leq \mathrm{eS} \cdot \mathrm{t} \leq 20$. Durch Striche wird die Skala graphisch in $\mathrm{eS} \cdot \mathrm{t}$ Teile geteilt.

Die Stufenhöhe für die Bewertungen ist immer gleich eins. Die Werte des Beispiels sind $\mathrm{Su} = 70$, $\mathrm{So} = 130$, $\mathrm{dS} = 10$ und $\mathrm{eS} = 2$, womit $\mathrm{t} = 6$ und $\mathrm{eS} \cdot \mathrm{t} = 12$. Die Skalengrenzen wurden so gewählt, dass 100 dem Durchschnitt der zugehörigen Ausprägungen der bewerteten Merkmale entspricht und Werte außerhalb der Skalengrenzen sehr selten vorkommen sollten. Diese Art der Skalierung der Bewertungen ist nicht immer sinnvoll und üblich.

Zweckmäßig ist es, eine Bewertungsskala mit nicht zu vielen, aber auch nicht zu wenigen Stufen zu wählen. Bei vielen Stufen ist es für die Beobachter schwierig zu entscheiden, ob sie z.B. mit 86 oder 87 bewerten sollen, andererseits ist bei nur wenigen Stufen die Unsicherheit allein dadurch recht hoch. Empfohlen wird eine Skala mit etwa 8 bis 15 Stufen, z.B. $\mathrm{Su} = 1$, $\mathrm{So} = 9$, $\mathrm{dS} = 1$, $\mathrm{eS} = 1$ mit 9 Stufen. Das bietet den Beobachtern genügend viele Möglichkeiten, die minimalen und maximalen Bewertungen zu wählen und diese auch z.B. einfach durch Ankreuzen von Zahlen 1 bis 9 auf einem dementsprechend gestalteten Formular abzugeben.

4) Auch der Datenteil des vierten Blocks besitzt nur eine Datenzeile. Diese enthält die Werte der folgenden drei Parameter: die globale relative Unsicherheit aller Gewichte in Prozent (0), das Gewicht der Vorauswahl im Verhältnis zur aktuellen Auswahlaufgabe (0.5), sowie als String (maximal 11 Zeichen) die Benennung der Art der Alternativen (Teilnehmer). (In Klammern stehen wieder die Werte des Beispiels.) Eine globale relative Unsicherheit von etwa 10 % bis 20 % ist im Allgemeinen realistisch und sinnvoll. Das Gewicht der Vorauswahl ist gleich null zu setzen, wenn Vorauswahlergebnisse nicht vorhanden sind. Als Benennung der Art der Alternativen eignen sich z.B. „Bewerber“ bei Personaleinstellungen oder „Gebäude“, wenn ein Gebäude gekauft werden soll.

5) Der Datenteil des nächsten, fünften Blocks besitzt für jede Alternative eine Datenzeile. Diese enthält aufsteigend die Nummer der Alternative, das Ergebnis der Vorauswahl, die Unsicherheit der Vorauswahl, sowie den Namen der Alternative (maximal 10 Zeichen). Auch wenn das Gewicht der Vorauswahl gleich null ist, Vorauswahlergebnisse also nicht vorhanden sind, müssen Werte für das Ergebnis und die Unsicherheit der Vorauswahl angegeben werden. Diese Werte sind jedoch beliebig und bedeutungslos. Die im Beispiel angegebenen Vorauswahlergebnisse wurden fiktiv so gewählt, dass in der Ergebnisausgabe von Teilnehmer 4 die Warnung „Gesamt- und Vorauswahlergebnis unterscheiden sich sehr stark“ erscheint. Die Namen der Teilnehmer des AC (Alternativen) sind Pseudonyme.

6) Der Datenteil des sechsten Blocks umfasst für jeden Beobachter eine Datenzeile. Diese enthält aufsteigend die Nummer des Beobachters, für jede Übung die Kennzahl 1, wenn der Beobachter die Alternativen in dieser Übung bewertet hat, sonst 0 (in der derzeitigen Programmversion ist nur 1 zulässig ♠), sowie den Namen des Beobachters (maximal 10 Zeichen). Im Beispiel sind die Namen der Beobachter Pseudonyme.

7) Der Datenteil des folgenden siebten Blocks hat für jede Übung eine Datenzeile, jedoch nur, wenn mehr als eine Übung angewendet wird. In der Datenzeile stehen lediglich aufsteigend die Nummer und der Titel der Übung (maximal 40 Zeichen). Bei nur einer Übung ist der Datenteil leer.

8) Im achten Block besitzt der Datenteil für jede Merkmalsgruppe eine Datenzeile. In dieser stehen aufsteigend die Nummer der Gruppe, die Anzahl der zur Gruppe gehörenden Merkmale, das Gewicht und der Titel der Merkmalsgruppe (maximal 40 Zeichen). Im Beispiel wurde das Gewicht einer Gruppe gleich der Anzahl der zur Gruppe gehörenden Merkmale gesetzt, allen Merkmalen somit gleiches Gewicht gegeben. Das ist nicht immer sinnvoll.

9) Im neunten Block hat der Datenteil für jedes Merkmal eine Datenzeile. In dieser stehen aufsteigend die Nummer des Merkmals, die Nummer der Merkmalsgruppe, zu der das Merkmal gehört, das Gewicht des Merkmals innerhalb seiner Gruppe, ein Akzeptanzgrenzwert, der von den Ergebnissen der Alternativen nicht unterschritten werden sollte, die Gewichte der Übungen bezüglich des Merkmals, sowie der Titel des Merkmals (maximal 40 Zeichen). Im Beispiel wurde allen Merkmalen das gleiche Gewicht zugewiesen, ebenso allen Übungen bezüglich jedes Merkmals. Das entspricht zwar der tatsächlich so erfolgten konventionellen, nichtquantitativen Auswertung, die Ergebnisse zeigen aber, dass dies nicht sinnvoll sein kann (Abschnitt 3.7). Alle Akzeptanzgrenzwerte der Merkmale wurden fiktiv gleich 90 gesetzt.

Die Gewichte der Merkmale brauchen nur innerhalb einer Merkmalsgruppe im Verhältnis zueinander gesetzt zu werden (z.B. 2:5:7 bei drei Merkmalen). Das ist ein Vorteil der Gruppierung, weil dadurch nur wenige Merkmale miteinander verglichen werden müssen. Auch die Gewichte der Übungen brauchen nur hinsichtlich eines Merkmals im Verhältnis zueinander entsprechend ihrer Aussagekraft für dieses Merkmal angegeben zu werden. Ebenso genügt es, die Gewichte der Merkmalsgruppen im Verhältnis zueinander zu setzen. Die absolute Wichtung übernimmt das Programm. Die Gewichte der Merkmale oder Übungen können auch gleich null sein, aber jeweils nicht alle in einer Gruppe bzw. zu einem Merkmal. In diesem Fall ist das Gewicht der Gruppe bzw. des Merkmals gleich null zu setzen. Der Akzeptanzgrenzwert bleibt unbeachtet, wenn er kleiner oder gleich der unteren Grenze Su der Bewertungsskala gesetzt wird. Dadurch kann er gegebenenfalls leicht ausgeschaltet werden.

2.4 Angabe der Bewertungsdaten

Der Datenteil des letzten, zehnten Blocks der Eingabedatei ist komplizierter als die vorangehenden Datenteile aufgebaut. Er enthält die Einzelbewertungen, die von den Beobachtern zu den Merkmalen der Alternativen in den Übungen abgegeben worden sind. Zweckmäßiger wäre es, für diese Bewertungen eine eigene Datei vorzusehen ♠. Für die Praxis ist geplant, die Bewertungen auf geeignet gestalteten, auf dem Bildschirm dargestellten Formularen durch Anklicken einzugeben, die so erfassten Bewertungen in einer Datei abzulegen und von dort aus vom Auswahlprogramm QWAHL auszuwerten ♠.

Der Datenteil besteht aus einer Folge von Untereinheiten. Jede Untereinheit beginnt mit einer Leerzeile, gefolgt von einer Kopfzeile, die die Nummer einer Alternative und die Nummer eines Beobachters trägt. Danach folgt für jedes Merkmal eine Bewertungszeile für die einzelnen Bewertungen, die von dem Beobachter zu dem Merkmal

und zu der Alternative in den Übungen abgegeben worden sind. Die Bewertungzeile enthält aufsteigend die Nummer des Merkmals und zu jeder Übung ein Datenpaar, das jeweils die minimale und die maximale Bewertung darstellt (Abschnitt 1.2). Hat ein Beobachter an einer Übung nicht teilgenommen oder besitzt eine Übung für das jeweilige Merkmal keine Aussagekraft, ist also das Gewicht dieser Übung für das Merkmal gleich null, so müssen zwar auch Bewertungen angegeben werden, diese sind aber beliebig und bedeutungslos.

Manchmal liegt die greifbare Information zu den einzelnen Merkmalen a l l e r Alternativen schon vor, bevor die Beobachter sie hinsichtlich der Ausprägung der Merkmale beurteilen und ihre Bewertungen zu den Merkmalen abgeben. In diesem Fall können die Beobachter die sich entsprechenden Merkmale der Alternativen miteinander vergleichen. Wenn eine Entscheidung zwischen den Alternativen zu treffen ist, kann es dann zweckmäßig sein, dass die Beobachter zu jedem Merkmal der besten Alternative die maximale Bewertung gleich der oberen Skalengrenze So und der schlechtesten Alternative die minimale Bewertung gleich der unteren Skalengrenze Su erteilen. Das ist sinnvoll, weil bei Entscheidungen nur Bewertungsdifferenzen, nicht aber die Bewertungen selbst relevant sind. Verschiebung und Dehnung der Bewertungsskala sind bedeutungslos.

Das Programm prüft jedes Datenpaar a,b auf Zulässigkeit und korrigiert es gegebenenfalls wie folgt: Wenn a $>$ b, werden a und b zunächst miteinander vertauscht. Wenn danach festgestellt wird, dass b e i d e Daten außerhalb des durch Su und So eingegrenzten Bewertungsskalenbereichs liegen, werden a durch Su und b durch So ersetzt. Das entspricht dem Fall nicht vorhandener Information, also einer fehlenden Bewertung. Liegt nur ein Datum außerhalb dieses Bereichs, wird dieses Datum gleich dem anderen Datum innerhalb des Bereichs gesetzt. Das entspricht dem Fall maximaler Information, bei dem minimale und maximale Bewertung gleich sind, die Bewertung also genau ist. Das eine Datum außerhalb des Bereichs dient demnach als Ersatzwert und Anzeige dafür, dass nur eine genaue Bewertung angegeben ist. Die Korrektur gilt nur für die Berechnungen, die Eingabedatei selbst wird nicht verändert. Im Beispiel (Anlage 1) kommen im Datenteil des zehnten Blocks, der die Bewertungen enthält, häufig Datenpaare a,b vor, bei denen a $= 0$ und b zwischen Su $= 70$ und So $= 130$ liegt. In diesen Fällen wurde vom Beobachter nur die Bewertung b als sicher genannt. Der Wert a $= 0$ zeigt dies an, dient als Ersatzwert. Dementsprechend ersetzt das Programm a durch b. Es kommen aber auch Datenpaare mit a $= 0$ und b $= 0$ vor. In diesen Fällen hat der Beobachter gar keine Bewertung abgegeben, also keine Information erkannt. Das Programm korrigiert deshalb und ersetzt a durch Su

und b durch So. Die völlige Unsicherheit des Beobachters geht auf diese Weise in die Auswertung ein.

2.5 Fehlermeldungen

Wird beim Lesen der Eingabedatei ein Datenfehler festgestellt, meldet das Programm eine Fehlernummer, die Fehlerart, gibt einige weitere Hinweise dazu und beendet sich nach ein- oder zweimaligem Drücken einer beliebigen Taste.

Das Programm prüft zwar einige Daten auf Zulässigkeit und Plausibilität, u.a. die Daten zur Festlegung der Bewertungsskala und ob die Vorauswahlergebnisse zu dieser Skala passen, es kann aber die Sorgfalt des Anwenders nicht ersetzen. Insbesondere zählt das Programm Datenzeilen der Datenteile der Blöcke 5 bis 10 und vergleicht den laufenden Zählwert jeweils mit dem ersten Datum der gerade zu lesenden Datenzeile. Diese beiden Zahlen müssen übereinstimmen, wenn nicht, enthält die Fehlermeldung das Wort „Zählung“, einen Hinweis auf die Art der gerade zu lesenden Daten und die Angabe der beiden differierenden Zahlen. Ein solcher Zählfehler wird häufig durch ein Tabulatorzeichen in der Eingabedatei verursacht, weil manche Editorprogramme (wozu das Programm EDIT nicht gehört) beim Abspeichern einer bearbeiteten Datei längere Folgen von Zwischenraumzeichen mit Hilfe von Tabulatorzeichen verkürzen. Das kann u.U. durch Ändern des Abspeichermodus des Editorprogramms vermieden werden, z.B. beim IBM-Editor PE II durch den Befehl SET BLANKCOMPRESS OFF.

3 Bedienung, Programmfunktionen, Ergebnisausgabe

3.1 Allgemeines

Die Ausgabe der vom Programm QWAHL errechneten Ergebnisse zu den einzelnen Alternativen erfolgt mit den jeweils zugehörigen Unsicherheiten und Rängen in nummerischer und graphischer Darstellung auf den Bildschirm. Sie ist in Einzelbildern und Bildfolgen organisiert, die den unterschiedlichen Programmfunktionen und Fragestellungen entsprechen und zwischen denen direkt oder über ein Menü durch Drücken einer oder weniger Tasten leicht umgeschaltet werden kann. Die Ergebnisausgabe in zwar vereinfachter, aber vollständiger nummerischer Darstellung in eine Datei oder mittels eines Druckers ist ebenfalls möglich (Beispiel siehe Anlage 2).

Um die folgende Beschreibung der Bedienung und der einzelnen Funktionen besser zu verstehen, starte man das Programm mit der beigefügten Beispiel-Eingabedatei QWAHL.EIN (Abschnitt 1.4 und Anlage 1).

Nach dem Start, dem Lesen der Eingabedatei und allen Berechnungen erscheint ein Bild mit dem Vergleich der Gesamtergebnisse zu den Alternativen (Funktion 1, Abschnitt 3.3, Figur 1). Drücken der Eingabetaste (Enter-Taste) führt zum Menübild.

3.2 Menü und Hilfe, allgemeine Bedienung

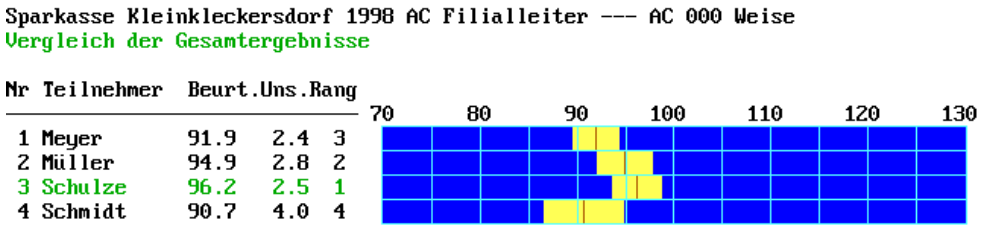
Das Menübild wird von jedem anderen Bild aus durch Drücken der Eingabetaste aufgerufen oder auch durch Drücken einer beliebigen anderen Taste, der im Folgenden nicht ausdrücklich eine andere Aufgabe zugewiesen wird.

Das Menübild enthält den Programmnamen QWAHL, die Versionsbezeichnung, eine kurze Bedienungsanleitung und das Menü mit Kurzerläuterung der einzelnen Funktionen.

Drücken der Buchstabentaste d leitet die nummerische Ausgabe der Ergebnisse aller Alternativen in eine ASCII-Datei oder mittels eines Druckers ein (Anlage 2). Das Programm fordert die Eingabe des Pfades und des Namens der Datei an. Wird daraufhin nur die Eingabetaste gedrückt, erfolgt der Ausdruck am Drucker LPT1. Drücken der Buchstabentaste e beendet das Programm.

Durch Drücken einer der Zifferntasten 1 bis 9, die den Funktionsnummern entsprechen, lässt sich vom Menübild auf jedes andere Bild umschalten. Nach Drücken einer der Tasten 5 bis 9 fordert das Programm die Eingabe einer Bildnummer an, je nach gewählter Funktion ist dies die Nummer der Merkmalsgruppe, des Merkmals, der Übung bzw. der Alternative, die gerade interessiert. Wird die Nummer 0 eingegeben, kehrt das Programm zum Menübild zurück. Bei unzulässiger Eingabe wird eine korrigierte Eingabe verlangt.

Auch von jedem vom Menübild verschiedenen Bild kann durch Drücken einer der Zifferntasten 1 bis 9 auf jede andere Funktion umgeschaltet werden. Die Eingabe einer Bildnummer wird dabei aber nicht verlangt, sondern diese wird gleich 1 gesetzt. Es wird also immer zum ersten Bild der gewählten Funktion umgeschaltet. Zum folgenden oder vorangehenden Bild derselben Funktion kann durch Drücken der Taste $+$ bzw. $-$ weiter- bzw. zurückgeschaltet werden, ebenso vom letzten Bild wieder zum ersten bzw. umgekehrt. Auch zwischen den Bildern der Funktionen 1 bis 4 kann auf diese Weise hin und her geschaltet werden, denn sie bilden zusammen eine eigene Bildfolge. Zwischen den Bildern der Funktionen 8 und 9 (Abschnitt 3.6) für die nummerische bzw. graphische Darstellung der Ergebnisse derselben Alternative kann durch Drücken der Taste # umgeschaltet werden. Jedes Bild trägt den Titel der Auswahlaufgabe, die Kennzeichnung der Funktion und bei den Funktionen 5 bis 9 die Bildnummer und einen dazugehörigen Titel.



Figur 1: Beispiel der Bildschirmausgabe bei Funktion 1: Vergleich der Gesamtergebnisse der Alternativen (bei den Funktionen 2, 3, 5, 6, 7 ähnlich)

3.3 Vergleich der Gesamt-, Vorauswahl- oder Endergebnisse

Das einzige Bild der Funktion 1 erscheint sofort nach Beendigung aller Berechnungen (Beispiel siehe Figur 1). Es kann auch von jedem anderen Bild durch Drücken der Zifferntaste 1 aufgerufen werden. Es bietet einen Vergleich der Gesamtbewertungen der einzelnen Alternativen, gewichtet zusammengefasst aus allen Einzelbewertungen der Merkmale durch die Beobachter in den verschiedenen Übungen. Die Gesamtbewertungen der Alternativen und die ihnen jeweils zugehörigen Unsicherheiten werden nummerisch und graphisch dargestellt. Auch die jeweils zugehörigen Ränge werden nummerisch angegeben. Die Angaben zum Rang 1 erscheinen grün hervorgehoben.

Die graphische Darstellung ist allgemein wie folgt aufgebaut (Figuren 1 und 4): Unter der in der Eingabedatei definierten Bewertungsskala erscheinen in den Angabezeilen dunkelblaue waagerechte Balken (bei Funktion 9 teilweise auch grün). Ein hellblaues Gitter unterteilt die Balken in eS · t Teile und grenzt sie voneinander ab. Auf den Balken sind die Ergebnisse durch braune, bei den Merkmalen die Akzeptanzgrenzwerte durch lila senkrechte Striche markiert, die Unsicherheitsbereiche durch waagerechte gelbe Balken. Ein senkrechter brauner Ergebnisstrich steht jeweils in der Mitte des zugehörigen Unsicherheitsbalkens. Ein waagerechter lila Strich über den ganzen Balken zeigt eine sehr große Unsicherheit an, einen Unsicherheitsbereich länger als die halbe Skalenlänge. Zur Interpretation der Unsicherheitsbereiche siehe Abschnitt 3.7.

Die jeweils einzigen Bilder der Funktionen 2 und 3 erscheinen durch Drücken der Zifferntasten 2 bzw. 3. Die Funktionen 2 und 3 sind nur dann vorhanden, wenn in der Eingabedatei Vorauswahlergebnisse der Alternativen mit einem Gewicht größer als null angegeben werden. Die Funktion 2 bietet einen nummerischen und graphischen Vergleich dieser Vorauswahlergebnisse mit den zugehörigen Unsicherheiten und Rängen, die Funktion 3 einen entsprechenden Vergleich der Endergebnisse, d.h. der zusammen-

Sparkasse Kleinkleckersdorf 1998 AC Filialleiter --- AC 000 Weise
Vergleich der Ränge

Nr	Gruppe	Merkmal	Übung	Ränge: Teilnehmer Nr 1	2	3	4
1	Fachliche Kompetenzen			4	2	1	3
	1	Produktkenntnisse		3	1	2	4
	2	Fachkenntnisse		4	2	1	3
	3	Konzeptionelle Fähigkeiten		4	3	1	2
2	Soziale Kompetenzen			3	2	1	4
	4	Gesprächsführung	Führungskompetenz	4	2	1	3
	5	Überzeugungskraft	Durchsetzungsvermögen	3	2	1	4
	6	Kontaktfähigkeit	Kundenwirksamkeit	2	3	1	3
	7	Akquisitions-	Verhandlungsvermögen	3	2	1	4
	8	Konfliktfähigkeit		3	1	2	4
3	Persönliche Kompetenzen			3	2	1	4
	9	Lernfähigkeit		4	1	2	3
	10	Auftreten		3	2	1	4
	11	Belastbarkeit	Frustrationstoleranz	3	1	2	4
1	Mitarbeitergespräch			3	2	1	4
2	Kundengespräch			3	1	2	4
3	Konzeptentwicklung			4	2	1	3
	Gesamtergebnis			3	2	1	4
	Vorauswahlergebnis			4	2	3	1
	Endergebnis			4	2	3	1

Figur 2: Beispiel der Bildschirmausgabe bei Funktion 4: Vergleich der Ränge

gefassten Gesamt- und Vorauswahlergebnisse. Die Art der Darstellung ist dieselbe wie bei Funktion 1. Die jeweils einzigen Bilder der Funktionen 1 bis 4 bilden zusammen eine Bildfolge, zwischen denen auch durch Drücken der Tasten + und − hin und her geschaltet werden kann (Abschnitt 3.2).

3.4 Vergleich der Ränge

Das einzige Bild der Funktion 4 wird durch Drücken der Zifferntaste 4 aufgerufen (Beispiel siehe Figur 2, siehe auch Abschnitt 3.3, letzter Absatz). Dieses Bild zeigt in Form einer Tabelle einen nummerischen Gesamtvergleich der Ränge der Alternativen in allen Merkmalsgruppen (blau), Merkmale (schwarz) und Übungen (Text grün, Ränge schwarz) sowie bei den Gesamt- (blau), Vorauswahl- (schwarz) und Endergebnissen (blau). Rang 1 erscheint jeweils grün hervorgehoben. Bei den Merkmalen bedeutet eine lila Rangangabe, auch bei Rang 1, dass das Ergebnis unter dem Akzeptanzgrenzwert liegt. Die Rangfolgen werden nach der Größe der sich entsprechenden Ergebnisse der Alternativen gebildet. Zwei oder mehrere gleiche Ergebnisse erhalten den gleichen Rang. Dann fehlen dementsprechend viele folgende Ränge. Im Beispiel (Figuren 2 und 3) ist der Rang 1 des Endergebnisses von Teilnehmer 4 im Vergleich zu den anderen Rängen nicht plausibel. Das liegt an dem offenbar zu groß angesetzten Gewicht 0,5 der Vorauswahl.

Sparkasse Kleinkleckersdorf 1998 AC Filialleiter --- AC 000 Weise
Ergebnisse Teilnehmer 4: Schmidt

Nr	Gruppe Merkmal Übung	Beurt.	Uns.	Rang
1	Fachliche Kompetenzen	96.1	9.2	3
1	Produktkenntnisse	92.9	13.8	4
2	Fachkenntnisse	98.0	17.5	3
3	Konzeptionelle Fähigkeiten	97.3	16.3	2
2	Soziale Kompetenzen	88.2	5.3	4
4	Gesprächsführung Führungskompetenz	89.4	10.9	3
5	Überzeugungskraft Durchsetzungsvermögen	88.1	12.2	4
6	Kontaktfähigkeit Kundenwirksamkeit	89.8	10.8	3
7	Akquisitions- Verhandlungsvermögen	86.8	12.5	4
8	Konfliktfähigkeit	87.0	12.6	4
3	Persönliche Kompetenzen	89.4	7.5	4
9	Lernfähigkeit	93.8	14.9	3
10	Auftreten	87.2	12.1	4
11	Belastbarkeit Frustrationstoleranz	87.2	11.5	4
1	Mitarbeitergespräch	90.7	2.4	4
2	Kundengespräch	89.1	4.6	4
3	Konzeptentwicklung	92.3	4.5	3
	Gesamtergebnis	90.7	4.0	4
	Vorauswahlergebnis	111.2	1.2	1
	Endergebnis	97.5	2.7	1

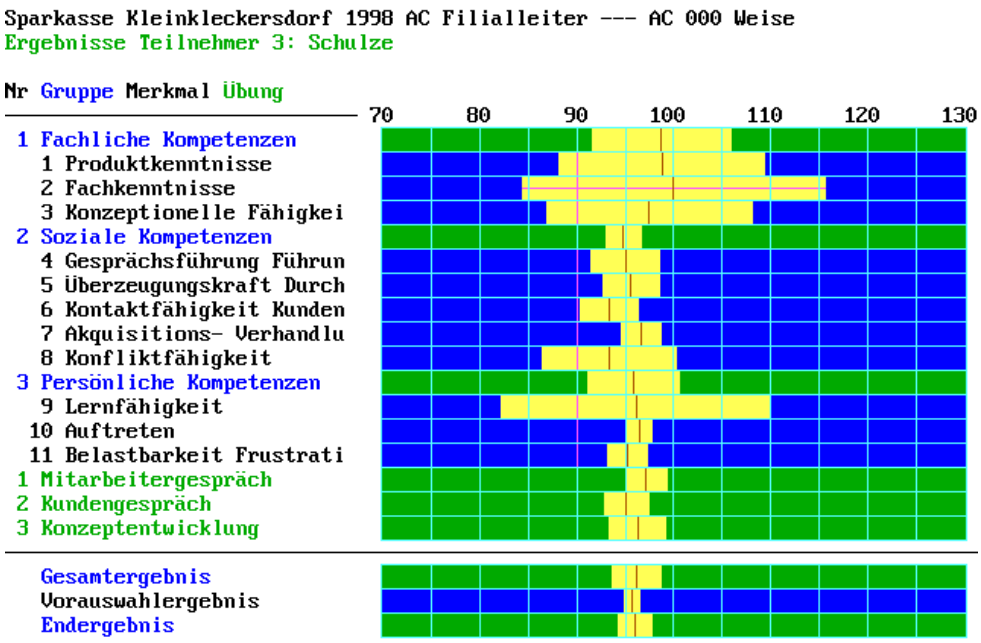
Gesamt- und Vorauswahlergebnis unterscheiden sich sehr stark

Figur 3: Beispiel der Bildschirmausgabe bei Funktion 8: Nummerische Darstellung aller Ergebnisse einer Alternative

3.5 Vergleich der Ergebnisse einer Merkmalsgruppe, eines Merkmals oder einer Übung

Die Bildfolgen der Funktionen 5, 6 und 7 werden durch Drücken der Zifferntasten 5, 6 bzw. 7 aufgerufen. Beim Aufruf vom Menübild her muss noch die gewünschte Bildnummer der Folge nach Aufforderung eingegeben werden. Die Bildnummer ist identisch mit der Nummer der Merkmalsgruppe (Funktion 5), des Merkmals (Funktion 6) oder Übung (Funktion 7). Beim Aufruf von einem anderen Bild her erscheint immer das erste Bild der gewählten Funktion. Durch Drücken der Tasten + und − kann in der Bildfolge weiter- bzw. zurückgeschaltet werden (Abschnitt 3.2).

Die Bilder der Funktionen 5, 6 und 7 bieten einen nummerischen und graphischen Vergleich der Ergebnisse der Alternativen mit den zugehörigen Unsicherheiten und Rängen jeweils hinsichtlich einer Merkmalsgruppe, eines Merkmals bzw. einer Übung. Die Art der Darstellung ist dieselbe wie bei Funktion 1 (Abschnitt 3.3, Figur 1). Bei den Merkmalen (Funktion 6) bedeutet eine lila Ergebnisangabe, dass das Ergebnis unter dem Akzeptanzgrenzwert liegt. In der Graphik ist der Akzeptanzgrenzwert wie in Figur 4 durch einen senkrechten lila Strich markiert.



Figur 4: Beispiel der Bildschirmausgabe bei Funktion 9: Graphische Darstellung aller Ergebnisse einer Alternative

3.6 Aufstellung der Ergebnisse einer Alternative

Die Bildfolgen der Funktionen 8 und 9 werden durch Drücken der Zifferntasten 8 bzw. 9 aufgerufen (Beispiele siehe Figuren 3 und 4). Beim Aufruf vom Menübild her muss noch die gewünschte Bildnummer der Folge nach Aufforderung eingegeben werden. Die Bildnummer ist identisch mit der Nummer der Alternative. Beim Aufruf von einem anderen Bild her erscheint immer das erste Bild der gewählten Funktion. Durch Drücken der Tasten + und − kann in der Bildfolge weiter- bzw. zurückgeschaltet werden. Durch Drücken der Taste # lässt sich zwischen den Bildern gleicher Nummer der Funktionen 8 und 9 umschalten (Abschnitt 3.2).

Die Bilder der Funktionen 8 und 9 zeigen in nummerischer bzw. graphischer Darstellung alle Ergebnisse jeweils einer Alternative mit den zugehörigen Unsicherheiten und Rängen, letztere nur bei Funktion 8. Die Art der Darstellung ist ähnlich der der anderen Funktionen. Der Hintergrund der waagerechten Balken der Graphik ist der Übersichtlichkeit halber bei den Merkmalsgruppen, Übungen, Gesamt- und Endergebnissen grün, sonst blau gewählt. Falls sich die Gesamt- und Vorauswahlergebnisse so stark unterscheiden, dass die Unsicherheitsbereiche nicht überlappen, erfolgt ein roter Warnhinweis (Figur 3). Die Datei- und Druckausgabe der Ergebnisse ähnelt der Bildfolge der Funktion 8, ist aber farblos (siehe Anlage 2).

3.7 Interpretation der Ergebnisse und Unsicherheiten

Wichtig ist es, die berechneten Ergebnisse und die mit ihnen verbundenen Unsicherheiten richtig zu interpretieren. Siehe dazu die begleitende wissenschaftliche Arbeit, die die theoretische Grundlage des Programms QWAHL bildet. Gemäß der in der Eingabedatei gegebenen Information und der Bewertungsskala sind die Ergebnisse selbst die besten Schätzwerte der Eigenschaften (Messgrößen) der Alternativen, die den Merkmalen unterliegen oder zusammenfassend auch den Merkmalsgruppen, Gesamt-, Vorauswahl- und Endergebnissen zugeordnet werden. Die Ergebnisse x sind die Erwartungswerte und die zugehörigen Unsicherheiten u(x) sind die Standardabweichungen von Wahrscheinlichkeitsverteilungen, die nichts als die gerade vorliegende unvollständige Information zu den Eigenschaften wiedergeben. Diese Verteilungen sind k e i n e Verteilungen von Werten, wie sie bei wiederholten Versuchen zufällig auftreten, sondern basieren auf dem Prinzip, gleichartigen Möglichkeiten die gleiche Wahrscheinlichkeit zuzuweisen, wie die Wahrscheinlichkeit 1/6 für jede Seite eines Würfels, o h n e dass dieser jedoch wirklich geworfen wird. Die Unsicherheit zu einem berechneten Ergebnis enthält sowohl 1) die Unsicherheiten zu den Einzelbewertungen der beitragenden Merkmale der Alternativen durch die Beobachter, die durch die Angabe jeweils minimaler und maximaler Bewertungen ausgedrückt werden, als auch 2) die Streuung der Bewertungen durch die einzelnen Beobachter in den einzelnen Übungen, sowie 3) global die Unsicherheiten der Gewichte und außerdem 4) den Unsicherheitsbeitrag der Skalenstufung.

Die in den Graphiken gelb dargestellten Unsicherheitsbereiche haben die Grenzen $x \pm u(x)$. Diese Unsicherheitsbereiche sind keine Vertrauensbereiche der konventionellen Statistik, sondern die Bereiche derjenigen Werte, die aufgrund der vorliegenden Information den jeweiligen Merkmalen als vernünftige Schätzwerte zugewiesen werden können (Guide to the expression of uncertainty in measurement. International Organization for Standardization, Genf 1995). Die Unsicherheitsbereiche zu den Gesamt-, Vorauswahl- und Endergebnissen können allerdings näherungsweise als Vertrauensbereiche zur Wahrscheinlichkeit 0,68 ($\approx 2/3$) aufgefasst werden, weil die hier zugrunde liegenden Wahrscheinlichkeitsverteilungen oft näherungsweise Normalverteilungen sind. Die Ergebnisse für zwei Alternativen zu demselben Merkmal sind nur dann signifikant verschieden, wenn die beiden zugehörigen Unsicherheitsbereiche nicht überlappen. Es wird besonders darauf hingewiesen, dass die Unsicherheit eines Merkmals einer Alternative nicht bedeutet, dass das Merkmal in seiner Ausprägung schwankt, sondern nur, dass das Merkmal nicht genau bekannt ist. Die Unsicherheit ist dementsprechend ein Maß für die unvollständige Kenntnis eines Merkmals.

Die ausgedruckten Ergebnisse des Beispiels (Anlage 2) entsprechen im Wesentlichen denen der in der abschließenden Beobachterkonferenz des AC erfolgten konventionellen, nichtquantitativen Auswertung, auch die Unsicherheiten erscheinen realistisch und vernünftig. Allerdings gibt es Auffälligkeiten: Bei allen Alternativen weisen die Merkmale der Gruppe 1 sowie Merkmal 9 „Lernfähigkeit“ besonders große Unsicherheiten auf, am größten sind diese bei Merkmal 2 „Fachkenntnisse“. Das ist ein Zeichen dafür, dass die Übungen für diese Merkmale nicht gut geeignet sind. Bei Merkmal 2 kommen in Anlage 1 auch viele nicht abgegebene Bewertungen vor, d.h. Datenpaare $a = 0$, $b = 0$ (Abschnitt 2.4). Die Gewichte der Merkmalsgruppe 1 und des Merkmals 9 sollten reduziert oder sogar gleich null gesetzt werden. Auf der anderen Seite weisen herausragende Unsicherheiten bei einzelnen Alternativen auf spezifischen Informationsmangel hin, z.B. bei Teilnehmer 2 zu Merkmal 8 „Konfliktfähigkeit“ und bei Teilnehmer 4 zu allen Merkmalen. Offenbar hat sich Teilnehmer 4 im AC schwer einschätzbar verhalten. Das Gesamtergebnis des Teilnehmers 3 auf Rang 1 ist zwar besser als das des Teilnehmers 2 auf Rang 2, aber nicht signifikant besser, weil sich die diesen Ergebnissen zugeordneten Unsicherheitsbereiche überlappen. Die Endergebnisse dürfen hier nicht zum Vergleich herangezogen werden, weil die Daten der Vorauswahl fiktiv gesetzt worden sind.

Je mehr Merkmale (auch Übungen, Beobachter, Bewertungen) zum Ergebnis einer Merkmalsgruppe oder zum Gesamtergebnis beitragen, desto kleiner wird in der Regel die Unsicherheit zu diesem Ergebnis. Leider gilt hierbei die Faustregel, dass eine Halbierung der Unsicherheit die vierfache Anzahl an Merkmalen und einen dementsprechend höheren Aufwand erfordert. Ein Unsicherheitsbereich liegt immer innerhalb der Skalengrenzen und das zugehörige Ergebnis in der Mitte des Unsicherheitsbereichs. Bei großer Unsicherheit tendiert das Ergebnis deshalb zur Mitte der Skala. Dies ist bei kritischer Betrachtung der Ergebnisse zu beachten, insbesondere beim Vergleich zu anderen, zwar schlechteren Ergebnissen, zu denen aber kleinere Unsicherheiten gehören. Siehe hierzu in Anlage 2 die Ergebnisse der Alternativen zu Merkmal 2. Bei einem Merkmal deutet eine Unsicherheit $u(x) > 0{,}3\,(So - Su)$, was einem Unsicherheitsbereich länger als etwa 60 % der Skalenlänge entspricht, auf sich erheblich widersprechende Bewertungen durch die Beobachter hin, was ebenfalls bedeuten kann, dass die Übungen für das betrachtete Merkmal nicht aussagefähig genug sind. In diesem Fall ist die Unsicherheit größer als in dem Fall, dass gar keine Bewertungen vorliegen. Ein waagerechter lila Warnstrich auf einem Unsicherheitsbalken der Graphik zeigt an, dass der Unsicherheitsbereich länger als 50 % der Skalenlänge ist, d.h. ein kaum tragbarer Informationsmangel vorliegt (Beispiel siehe bei Merkmal 2 „Fachkenntnisse“ in Figur 4). Eine genauere Informationsbedingung aus der begleitenden Untersuchung ist noch zu ergänzen ♠ .

Anlage 1: Beispiel-Eingabedatei

```
QWAHL : Beispiel-Eingabedatei des Auswahlprogramms QWAHL
=====

Vor dem Start des Programms QWAHL ist der aktuelle Dateiname
in "QWAHL.EIN" zu ändern.
Jeder Datenteil wird durch eine vorausgehende, mit "*"
beginnende Kommentarzeile angekündigt. Davor darf beliebiger
Kommentar stehen.
Die Anzahl der Leerzeilen zwischen den Datenzeilen darf
nicht verändert werden (siehe z.B. unter 10).
Die Daten sind durch Zwischenraum oder "," zu trennen.
Als Dezimalzeichen dient "." (siehe z.B. unter 5).
Die Strings in den Datenzeilen dürfen kein Komma enthalten.

* 1) Titel der Auswahlaufgabe:
Sparkasse Kleinkleckersdorf AC Filialleiter --- AC 000 Weise

  2) Anzahlen: Alternativen, Beobachter, Übungen,
*              Merkmalsgruppen, Merkmale
4  3  3  3  11

  3) Bewertungsskala:
Bewertungsschrittweite ist immer 1
Su, So  untere und obere Grenze der Bewertungsskala
dS      Schrittweite der Skalenbeschriftung
eS      Skalenteiler der Schrittweite
Darstellungstechnische Bedingungen:
  Alle Skalenangaben ganzzahlig,
  -100 < Su < So < 1000 (mit Vorzeichen höchstens dreiziffrig),
  dS > 0, So - Su = dS * t (dS teilt So - Su).
  Für t sind nur die Werte 1 bis 8, 10 und 12 zulässig
  eS > 0, eS Teiler von dS,
  zweckmäßig, aber nicht notwendig: eS <= 5, 10 <= eS * t <= 20
*
70  130  10  2

  4) Parameter: globale relative Unsicherheit der Gewichte in Prozent,
                Gewicht Vorauswahl im Verhältnis zur aktuellen Aufgabe,
*               Art der Alternativen (max. 11 Zeichen)
0  0.5  Teilnehmer

  5) Alternativen: Nr, Ergebnis d. Vorauswahl,
*                   Unsicherheit d. Vorauswahl, Name (max. 10 Zeichen)
1    91.3  2.7  Meyer
2   100.7  3.5  Müller
3    95.7  0.8  Schulze
4   111.2  1.2  Schmidt

  6) Beobachter: Nr, Teilnahmen an d. Übungen (1 = ja, 0 = nein,
*                 derzeit nur 1 zulässig), Name (max. 10 Zeichen)
1    1  1  1    Hinz
2    1  1  1    Kunz
3    1  1  1    Schuster

  7) Übungen (nur, wenn Anzahl > 1): Nr, Titel (max. 40 Zeichen)
*    Diese Kommentarzeile muss auch erhalten bleiben, wenn Anzahl = 1
1           Mitarbeitergespräch
2           Kundengespräch
3           Konzeptentwicklung

  8) Merkmalsgruppen: Nr, Anzahl der Merkmale,
*                     Gewicht, Titel (max. 40 Zeichen)
```

```
1   3   3   Fachliche Kompetenzen
2   5   5   Soziale Kompetenzen
3   3   3   Persönliche Kompetenzen

  9) Merkmale: Nr, zu Gruppe, Gewicht innerhalb d. Gruppe,
               Akzeptanzgrenzwert,
*              Gewichte d. Übungen, Titel (max. 40 Zeichen)
1   1   1   90    1  1  1    Produktkenntnisse
2   1   1   90    1  1  1    Fachkenntnisse
3   1   1   90    1  1  1    Konzeptionelle Fähigkeiten
4   2   1   90    1  1  1    Gesprächsführung Führungskompetenz
5   2   1   90    1  1  1    Überzeugungskraft Durchsetzungsvermögen
6   2   1   90    1  1  1    Kontaktfähigkeit Kundenwirksamkeit
7   2   1   90    1  1  1    Akquisitions- Verhandlungsvermögen
8   2   1   90    1  1  1    Konfliktfähigkeit
9   3   1   90    1  1  1    Lernfähigkeit
10  3   1   90    1  1  1    Auftreten
11  3   1   90    1  1  1    Belastbarkeit Frustrationstoleranz

  10) Bewertungen:
1. Zeile jeder Dateneinheit: Leerzeile
2. Zeile:       Alternativen-Nr, Beobachter-Nr
weitere Zeilen: Merkmals-Nr, Bewertungen zum Merkmal
                in den einzelnen Übungen
           Jede Bewertung als Paar mit Min. und Max.
       0 0  bedeutet: keine Bewertung angegeben
       0 n  bedeutet: n ist allein als genaue Bewertung angegeben
*           (Min. = Max.). 0 n und n n sind daher identisch

1   1
1      0 97    0 91   0 0
2      0 0     0 0    0 0
3      0 100   0 91   0 0
4      0 85    0 0    0 92
5     85 105   0 94   0 95
6      0 90    0 95   0 97
7      0 92    0 94   0 94
8      0 88    0 91   0 88
9      0 96    0 95   0 0
10     0 94    0 96   0 100
11     0 96    0 96   0 92

1   2
1      0 97   89 91   0 0
2      0 98    0 91   0 0
3      0 99    0 0    0 89
4      0 82    0 89   0 88
5      0 89    0 91   0 81
6     91 95    0 89   0 83
7      0 92    0 89   0 81
8      0 88    0 89   0 81
9      0 81    0 91   0 89
10     0 89    0 91   0 81
11     0 99    0 91   0 91

1   3
1      0 92    0 92   0 0
2      0 88    0 89   0 0
3      0 94    0 92   0 92
4      0 82    0 85   0 85
5      0 82    0 91   0 91
6      0 93    0 91   0 92
7      0 89    0 90   0 90
8      0 91    0 91   0 91
```

```
9       0 0      0 92    0 91
10      0 86     0 90    0 91
11      0 89     0 91    0 86

2   1
1       0 101    0 101    0 0
2       0 0      0 0      0 0
3       0 0      0 0      0 0
4       0 93     0 96     0 94
5       0 93     0 95     0 92
6       0 91     0 90     0 89
7       0 90     0 91     0 90
8       0 0      0 0      0 89
9       0 0      0 0      0 0
10      0 92     0 94     0 89
11      0 99     0 100    0 95

2   2
1       0 100    0 100    0 0
2       0 0      0 0      0 0
3       0 91     0 92     0 0
4       0 91     0 94     0 89
5       0 89     0 95     0 89
6       0 85     0 89     0 89
7       0 85     0 89     0 89
8       0 93     0 95     0 92
9       0 0      0 91     0 0
10      0 88     0 91     0 91
11      0 96     0 96     0 94

2   3
1       0 97     0 98     0 0
2       0 98     0 97     0 0
3       0 95     0 97     0 95
4       0 92     0 92     0 93
5       0 93     0 94     0 92
6       0 91     0 92     0 92
7       0 94     0 95     0 93
8       0 95     0 0      0 92
9       0 0      0 0      0 0
10      0 94     0 95     0 96
11      0 95     0 96     0 97

3   1
1       0 100    0 97    0 0
2       0 0      0 0     0 0
3       0 0      0 0     0 0
4       0 99     0 95    0 92
5       0 95     0 96    0 98
6       0 95     0 95    0 98
7       0 98     0 98    0 96
8       0 91     0 0     0 92
9       0 0      0 0     0 0
10      0 96     0 98    0 97
11      0 97     0 94    0 98

3   2
1       0 101    0 91    0 0
2       0 0      0 0     0 0
3       0 97     0 91    0 94
4       0 98     0 89    0 93
5       0 99     0 91    0 92
6       0 90     0 88    0 94
7       0 97     0 92    0 95
```

```
8       0 92     0 91    0 98
9       0 92     0 88    0 92
10      0 95     0 97    0 96
11      0 97     0 95    0 96

3   3
1       0 100    0 100   0 0
2       0 100    0 100   0 0
3       0 100    0 97    0 98
4       0 99     0 92    0 98
5       0 100    0 95    0 94
6       0 91     0 94    0 95
7       0 99     0 98    0 97
8       0 94     0 94    0 88
9       0 0      0 93    0 0
10      0 98     0 97    0 94
11      0 96     0 92    0 92

4   1
1      90 94     0 80    0 0
2       0 0      0 0     0 0
3       0 0      0 0     0 0
4      93 95     0 81   85 90
5       0 91     0 70    0 82
6       0 92     0 78    0 90
7       0 91     0 72    0 75
8       0 91     0 70    0 75
9       0 90     0 0     0 0
10      0 90     0 72    0 78
11      0 92     0 78    0 78

4   2
1       0 92     0 89    0 0
2       0 0      0 0     0 0
3       0 87     0 89    0 0
4       0 87     0 82    0 82
5       0 87     0 82    0 89
6       0 87     0 82    0 87
7       0 87     0 82    0 83
8       0 88     0 82    0 89
9       0 84     0 85    0 85
10      0 88     0 84    0 82
11      0 87     0 82    0 82

4   3
1       0 83     0 0     0 0
2       0 82     0 0     0 0
3       0 0      0 0     0 0
4       0 91     0 0     0 0
5       0 92     0 0     0 0
6       0 92     0 0     0 0
7       0 91     0 0     0 0
8       0 88     0 0     0 0
9       0 0      0 0     0 0
10      0 91     0 0     0 0
11      0 86     0 0     0 0

Hier darf beliebiger Kommentar folgen
```

Anlage 2: Ausgabe zur Beispiel-Eingabedatei

Sparkasse Kleinkleckersdorf AC Filialleiter --- AC 000 Weise
Ergebnisse Teilnehmer 1: Meyer

Nr	Gruppe	Merkmal Übung	Bewert.	Uns.	Rang
1		Fachliche Kompetenzen	95.6	6.7	4
	1	Produktkenntnisse	95.4	10.9	3
	2	Fachkenntnisse	96.2	14.0	4
	3	Konzeptionelle Fähigkeiten	95.2	9.3	4
2		Soziale Kompetenzen	89.5	2.3	3
	4	Gesprächsführung Führungskompetenz	87.6*	8.0	4
	5	Überzeugungskraft Durchsetzungsvermögen	89.9*	5.3	3
	6	Kontaktfähigkeit Kundenwirksamkeit	91.4	3.8	2
	7	Akquisitions- Verhandlungsvermögen	90.1	3.7	3
	8	Konfliktfähigkeit	88.7*	3.0	3
3		Persönliche Kompetenzen	92.0	4.0	3
	9	Lernfähigkeit	92.8	10.0	4
	10	Auftreten	90.9	5.2	3
	11	Belastbarkeit Frustrationstoleranz	92.3	3.8	3
1		Mitarbeitergespräch	91.7	1.8	3
2		Kundengespräch	92.1	1.9	3
3		Konzeptentwicklung	91.8	3.0	4
		Gesamtergebnis	91.9	2.4	3
		Vorauswahlergebnis	91.3	2.7	4
		Endergebnis	91.7	1.8	4

* Akzeptanzgrenzwert unterschritten

Sparkasse Kleinkleckersdorf AC Filialleiter --- AC 000 Weise
Ergebnisse Teilnehmer 2: Müller

Nr	Gruppe	Merkmal Übung	Bewert.	Uns.	Rang
1		Fachliche Kompetenzen	98.6	7.4	2
	1	Produktkenntnisse	99.7	10.2	1
	2	Fachkenntnisse	99.4	15.6	2
	3	Konzeptionelle Fähigkeiten	96.7	12.2	3
2		Soziale Kompetenzen	92.1	2.4	2
	4	Gesprächsführung Führungskompetenz	92.7	1.9	2
	5	Überzeugungskraft Durchsetzungsvermögen	92.4	2.1	2
	6	Kontaktfähigkeit Kundenwirksamkeit	89.8*	2.1	3
	7	Akquisitions- Verhandlungsvermögen	90.7	2.9	2
	8	Konfliktfähigkeit	95.1	10.9	1
3		Persönliche Kompetenzen	95.9	5.7	2
	9	Lernfähigkeit	99.0	16.8	1
	10	Auftreten	92.2	2.6	2
	11	Belastbarkeit Frustrationstoleranz	96.4	1.9	1
1		Mitarbeitergespräch	94.6	2.5	2
2		Kundengespräch	95.6	2.5	1
3		Konzeptentwicklung	94.6	3.1	2
		Gesamtergebnis	94.9	2.8	2
		Vorauswahlergebnis	100.7	3.5	2
		Endergebnis	96.8	2.2	2

* Akzeptanzgrenzwert unterschritten

Sparkasse Kleinkleckersdorf AC Filialleiter --- AC 000 Weise
Ergebnisse Teilnehmer 3: Schulze

Nr	Gruppe	Merkmal Übung	Bewert.	Uns.	Rang
1		Fachliche Kompetenzen	98.7	7.2	1
	1	Produktkenntnisse	98.8	10.6	2
	2	Fachkenntnisse	100.0	15.5	1
	3	Konzeptionelle Fähigkeiten	97.4	10.6	1
2		Soziale Kompetenzen	94.8	1.8	1
	4	Gesprächsführung Führungskompetenz	95.0	3.5	1
	5	Überzeugungskraft Durchsetzungsvermögen	95.6	2.9	1
	6	Kontaktfähigkeit Kundenwirksamkeit	93.3	2.9	1
	7	Akquisitions- Verhandlungsvermögen	96.7	2.0	1
	8	Konfliktfähigkeit	93.3	6.8	2
3		Persönliche Kompetenzen	95.9	4.7	1
	9	Lernfähigkeit	96.1	13.9	2
	10	Auftreten	96.4	1.3	1
	11	Belastbarkeit Frustrationstoleranz	95.2	2.1	2
1		Mitarbeitergespräch	97.2	2.1	1
2		Kundengespräch	95.1	2.3	2
3		Konzeptentwicklung	96.3	2.9	1
		Gesamtergebnis	96.2	2.5	1
		Vorauswahlergebnis	95.7	0.8	3
		Endergebnis	96.0	1.7	3

Sparkasse Kleinkleckersdorf AC Filialleiter --- AC 000 Weise
Ergebnisse Teilnehmer 4: Schmidt

Nr	Gruppe	Merkmal Übung	Bewert.	Uns.	Rang
1		Fachliche Kompetenzen	96.1	9.2	3
	1	Produktkenntnisse	92.9	13.8	4
	2	Fachkenntnisse	98.0	17.5	3
	3	Konzeptionelle Fähigkeiten	97.3	16.3	2
2		Soziale Kompetenzen	88.2	5.3	4
	4	Gesprächsführung Führungskompetenz	89.4*	10.9	3
	5	Überzeugungskraft Durchsetzungsvermögen	88.1*	12.2	4
	6	Kontaktfähigkeit Kundenwirksamkeit	89.8*	10.8	3
	7	Akquisitions- Verhandlungsvermögen	86.8*	12.5	4
	8	Konfliktfähigkeit	87.0*	12.6	4
3		Persönliche Kompetenzen	89.4	7.5	4
	9	Lernfähigkeit	93.8	14.9	3
	10	Auftreten	87.2*	12.4	4
	11	Belastbarkeit Frustrationstoleranz	87.2*	11.5	4
1		Mitarbeitergespräch	90.7	2.4	4
2		Kundengespräch	89.1	4.6	4
3		Konzeptentwicklung	92.3	4.5	3
		Gesamtergebnis	90.7	4.0	4
		Vorauswahlergebnis	111.2	1.2	1
		Endergebnis	97.5	2.7	1

Gesamt- und Vorauswahlergebnis unterscheiden sich sehr stark
* Akzeptanzgrenzwert unterschritten